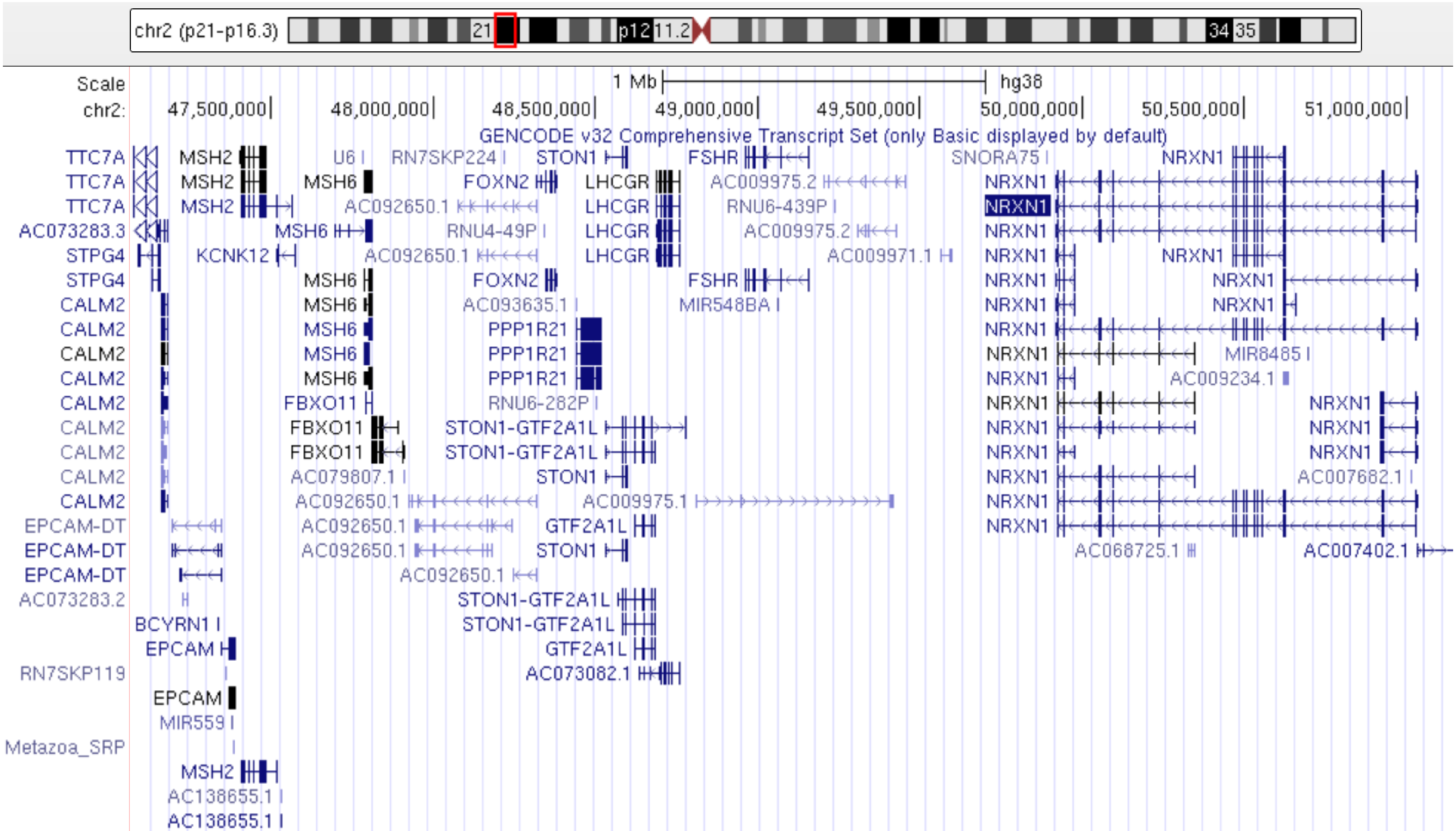
Mutations in Space:

Genes and Consequences

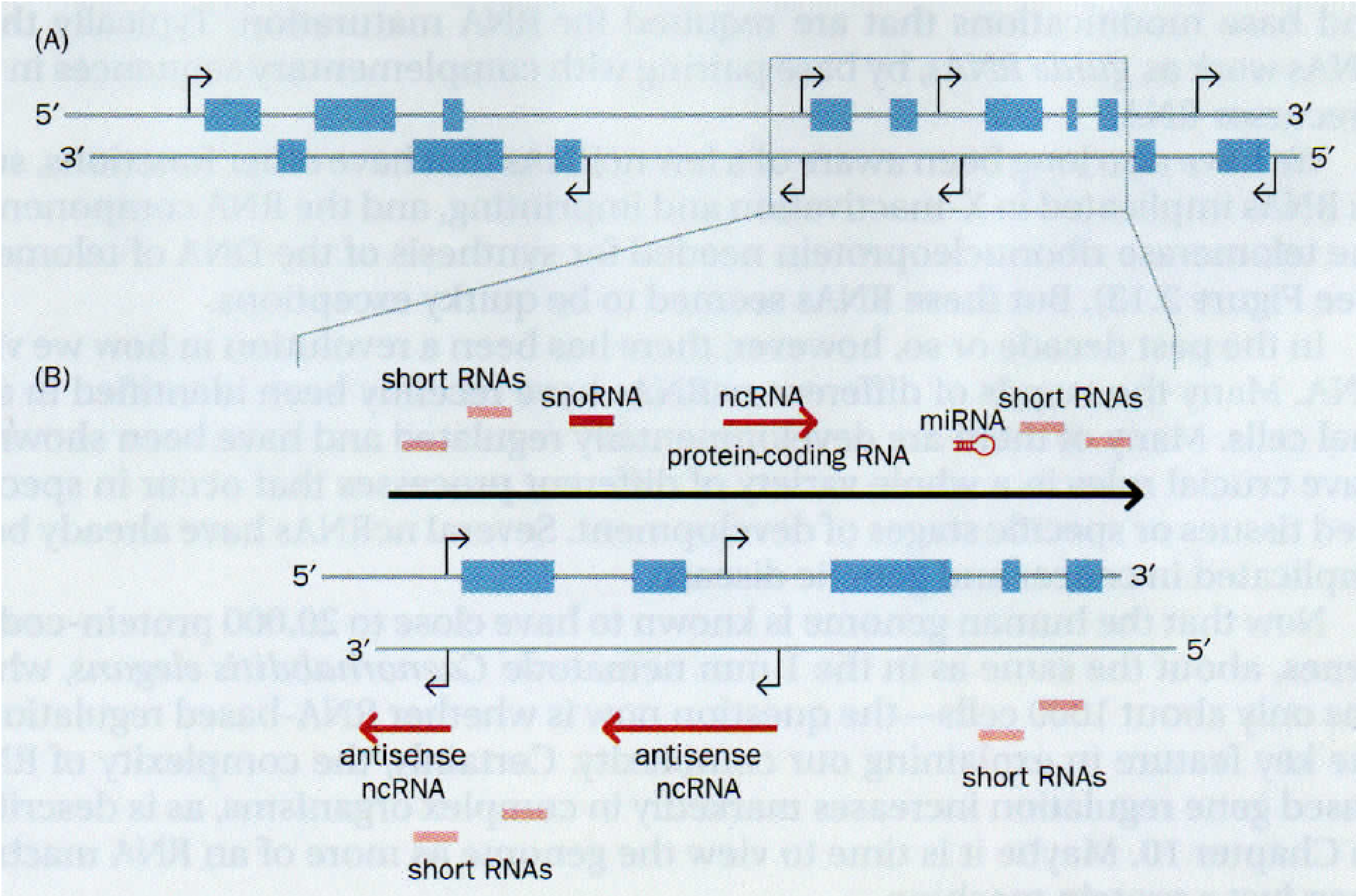
Lecture plan

- Overview of human genes structure and processing
- Alternative splicing
- Epigenetics. Chromosomal imprinting.
- Variant annotation. ENSEMBL Variant Effect Predictor: impact and consequences
- Protein-truncating and loss-of-function variants
- Missense variants, inframe indels
- Synonymous and regulatory variants
- Variant effect, dominant and recessive variants, gain- and loss-of-function

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

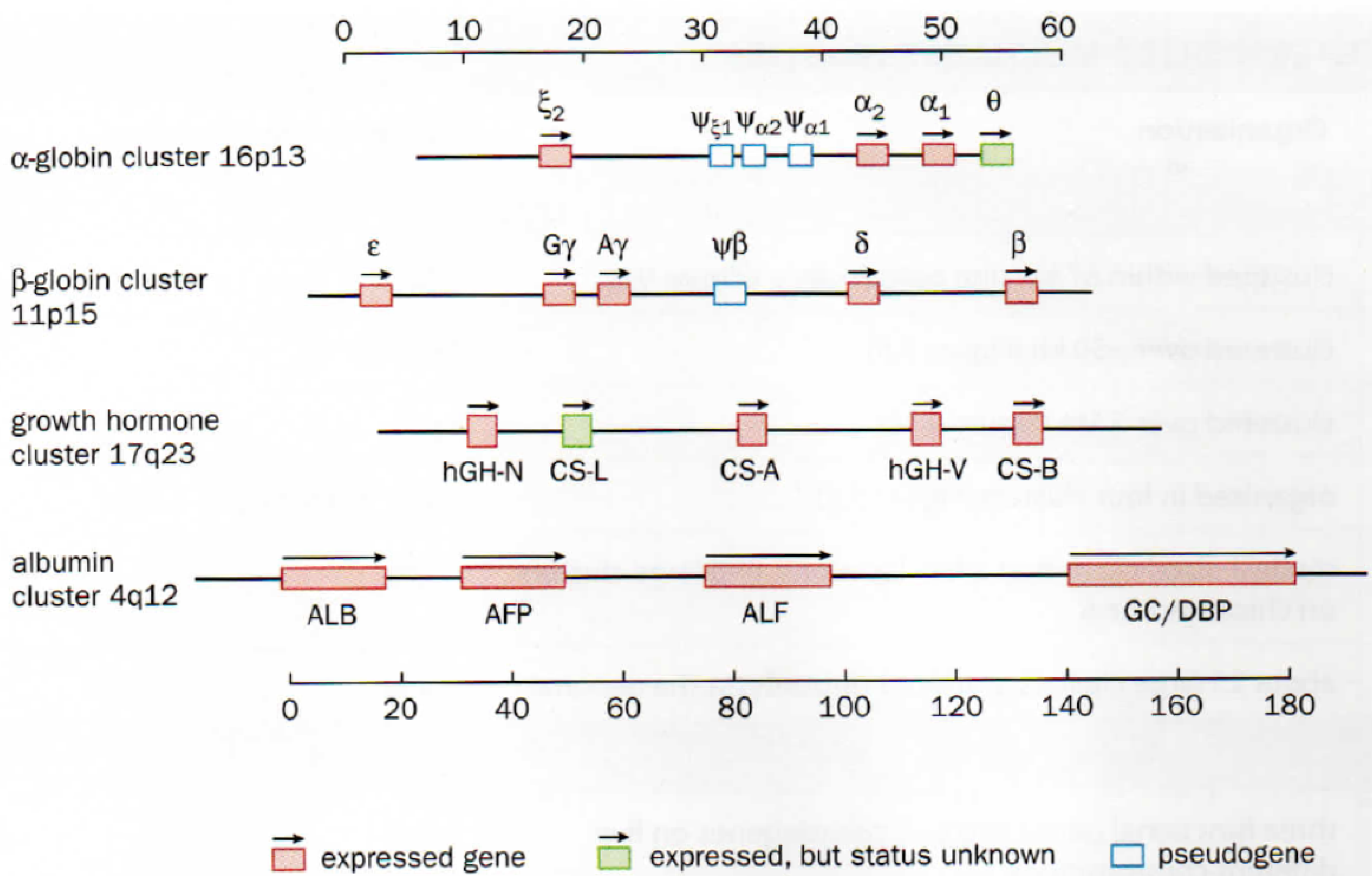


Blurring of gene boundaries



Strachan, Read – *Human Molecular Genetics*

Multigene families



Strachan, Read – *Human Molecular Genetics*

Multigene families

TABLE 9.6 EXAMPLES OF CLUSTERED AND INTERSPERSED MULTIGENE FAMILIES

Family	Copy no.	Organization	Chromosome location(s)
CLUSTERED GENE FAMILIES			
Growth hormone gene cluster	5	clustered within 67 kb; one pseudogene (Figure 9.8)	17q24
α-Globin gene cluster	7	clustered over ~50 kb (Figure 9.8)	16p13
Class I HLA heavy chain genes	~20	clustered over 2 Mb (Figure 9.10)	6p21
HOX genes	38	organized in four clusters (Figure 5.5)	2q31, 7p15, 12q13, 17q21
Histone gene family	61	modest-sized clusters at a few locations; two large clusters on chromosome 6	many
Olfactory receptor gene family	> 900	about 25 large clusters scattered throughout the genome	many
INTERSPERSED GENE FAMILIES			
Aldolase	5	three functional genes and two pseudogenes on five different chromosomes	many
PAX	9	all nine are functional genes	many
NF1 (neurofibromatosis type I)	> 12	one functional gene at 22q11; others are nonprocessed pseudogenes or gene fragments (Figure 9.11)	many, mostly pericentromeric
Ferritin heavy chain	20	one functional gene on chromosome 11; most are processed pseudogenes	many

Strachan, Read – *Human Molecular Genetics*

Human protein classes

PANTHER Protein Class

Total # Genes: 20996   Total # protein class hits: 11214

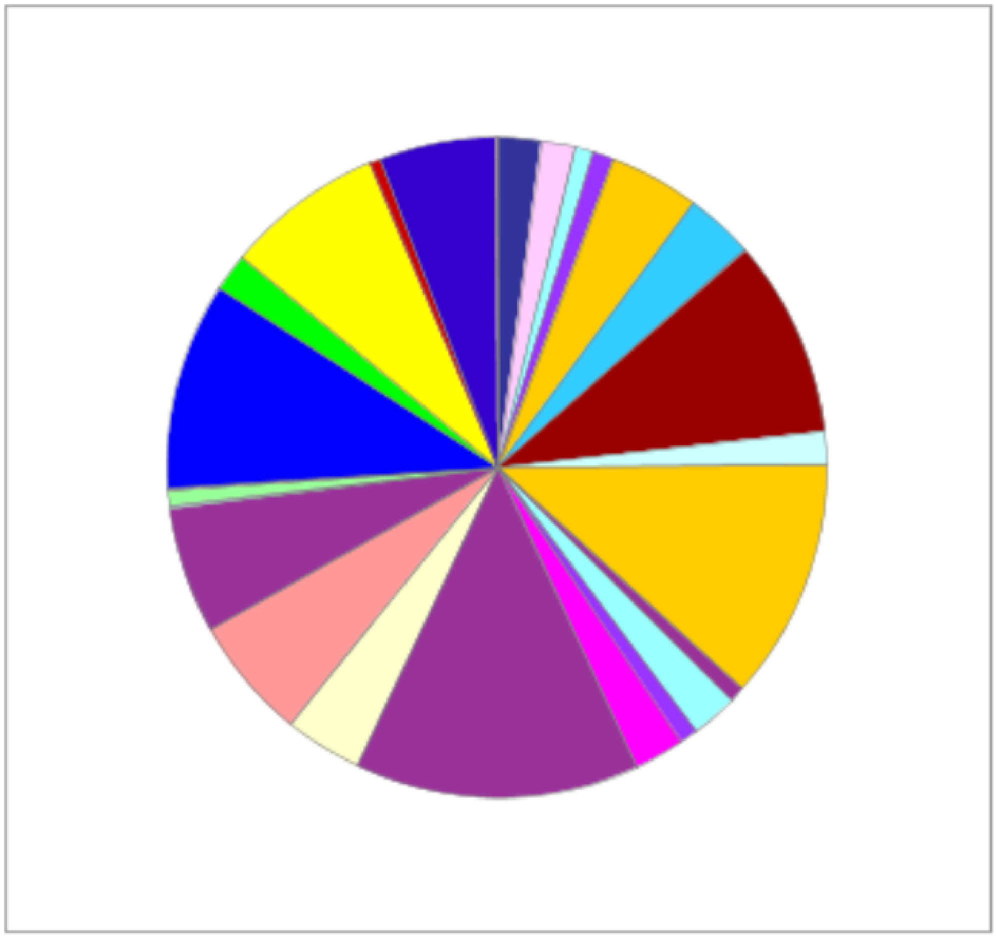


**Chart tooltips are read as: Category name (Accession): # genes; Percent of gene hit against total # genes; Percent of gene hit against total # Protein Class hits

Click to get gene list for a category:

- calcium-binding protein (PC00060)
- cell adhesion molecule (PC00069)
- cell junction protein (PC00070)
- chaperone (PC00072)
- cytoskeletal protein (PC00085)
- defense/immunity protein (PC00090)
- enzyme modulator (PC00095)
- extracellular matrix protein (PC00102)
- hydrolase (PC00121)
- isomerase (PC00135)
- ligase (PC00142)
- lyase (PC00144)
- membrane traffic protein (PC00150)
- nucleic acid binding (PC00171)
- oxidoreductase (PC00176)
- receptor (PC00197)
- signaling molecule (PC00207)
- storage protein (PC00210)
- structural protein (PC00211)
- surfactant (PC00212)
- transcription factor (PC00218)
- transfer/carrier protein (PC00219)
- transferase (PC00220)
- transmembrane receptor regulatory/adaptor protein (PC00226)
- transporter (PC00227)
- viral protein (PC00237)

Human protein classes

1	Nucleic acid binding (PC00171)	1567
2	Hydrolase (PC00121)	1322
3	Transcription factor (PC00218)	1138
4	Enzyme modulator (PC00095)	1079
5	Transferase (PC00220)	867
6	Signaling molecule (PC00207)	693
7	Receptor (PC00197)	675
8	Transporter (PC00227)	638
9	Cytoskeletal protein (PC00085)	497
10	Oxidoreductase (PC00176)	424
11	Defense/immunity protein (PC00090)	386
12	Membrane traffic protein (PC00150)	280
13	Ligase (PC00142)	250
14	Calcium-binding protein (PC00060)	237
15	Transfer/carrier protein (PC00219)	203
16	Cell adhesion molecule (PC00069)	195
17	Extracellular matrix protein (PC00102)	190
18	Chaperone (PC00072)	111
19	Cell junction protein (PC00070)	98
20	Lyase (PC00144)	97
21	Isomerase (PC00135)	85
22	Structural protein (PC00211)	84
23	Transmembrane receptor regulatory/adaptor protein (PC00226	64
24	Storage protein (PC00210)	18
25	Viral protein (PC00237)	8
26	Surfactant (PC00212)	8
27	Unknown	9782
	Total	20996

Exercise: think of appropriate questions

HGNC

HUGO Gene Nomenclature Committee

The resource for approved human gene nomenclature

UniProt

UniProtKB ▾

BLAST Align Retrieve/ID mapping Peptide search

GeneCards®: The Human Gene Database

GeneCards is a searchable, integrative database that provides comprehensive, user-friendly information on all annotated and predicted human genes. The knowledgebase automatically integrates gene-centric data from ~150 web sources, including genomic, transcriptomic, proteomic, genetic, clinical and functional information.

e! Ensembl BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog

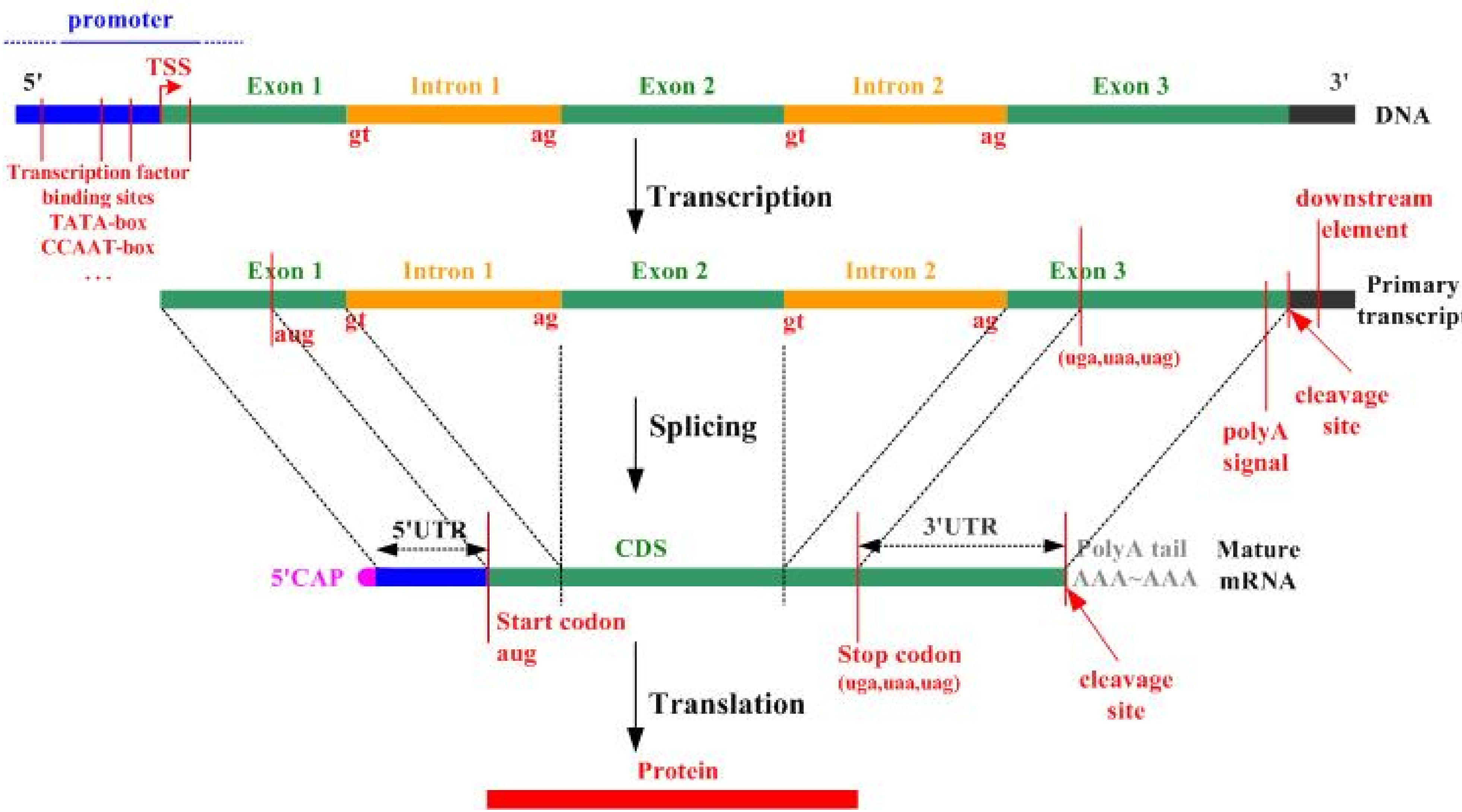
Human (GRCh38.p13) ▾

Search Human (*Homo sapiens*)

Search all categories ▾ | *Search Human...* | Go

e.g. **BRCA2** or **17:63992802-64038237** or **rs699** or **osteoarthritis**

Human gene structure and processing



Note: CDS (coding sequence) vs. mRNA, splicing sites, stop and start codons

Exercise: draw a typical human gene

Human gene structure and processing

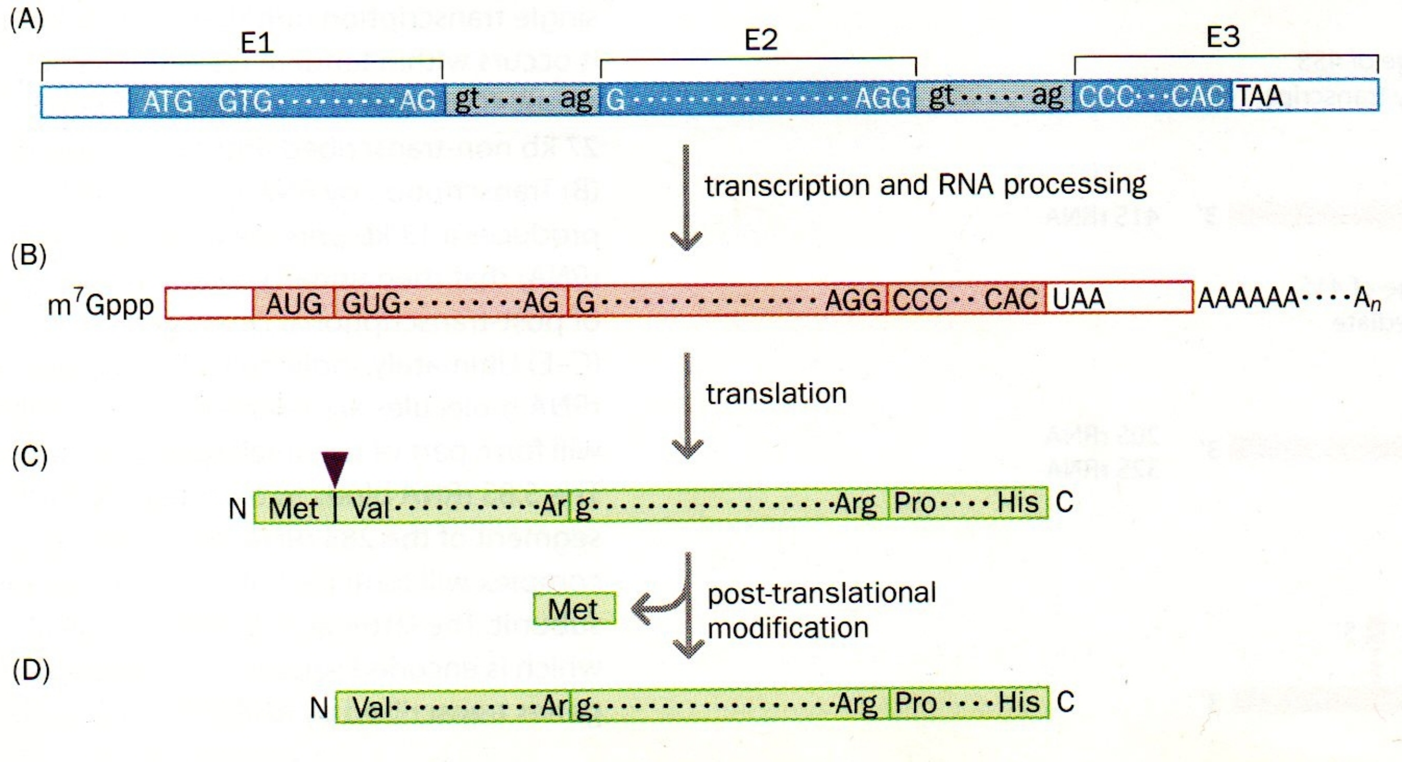


Figure 1.23 Transcription and translation of the human β-globin . (A) The β-globin gene comprises three exons (El-E3) and two introns. The 5'-end sequence of El and the 3' end sequence of E3 are noncoding sequences (unshaded sections). (B) These sequences are transcribed and so occur at the 5' and 3' ends (unshaded sections) of the β-globin mRNA that emerges from RNA processing. (C) Some codons can be specified by bases that are separated by an intron. The Arg104 is encoded by the last three nucleotides (AGG) of exon 2 but the Arg30 is encoded by an AGG codon whose first two bases are encoded by the last two nucleotides of exon 1 and whose third base is encoded by the first nucleotide of exon 2. (D) During post-translational modification the 147·amino acid precursor polypeptide undergoes cleavage to remove ils *N*-terminal methionine residue, to generate the mature 146-residue β-globin protein. The flanking *N* and *C* symbols to the left and right, respectively, in (C) and (D) depict the *N*-terminus and *C*-terminus.

Strachan, Read – *Human Molecular Genetics*

Human gene structure and processing

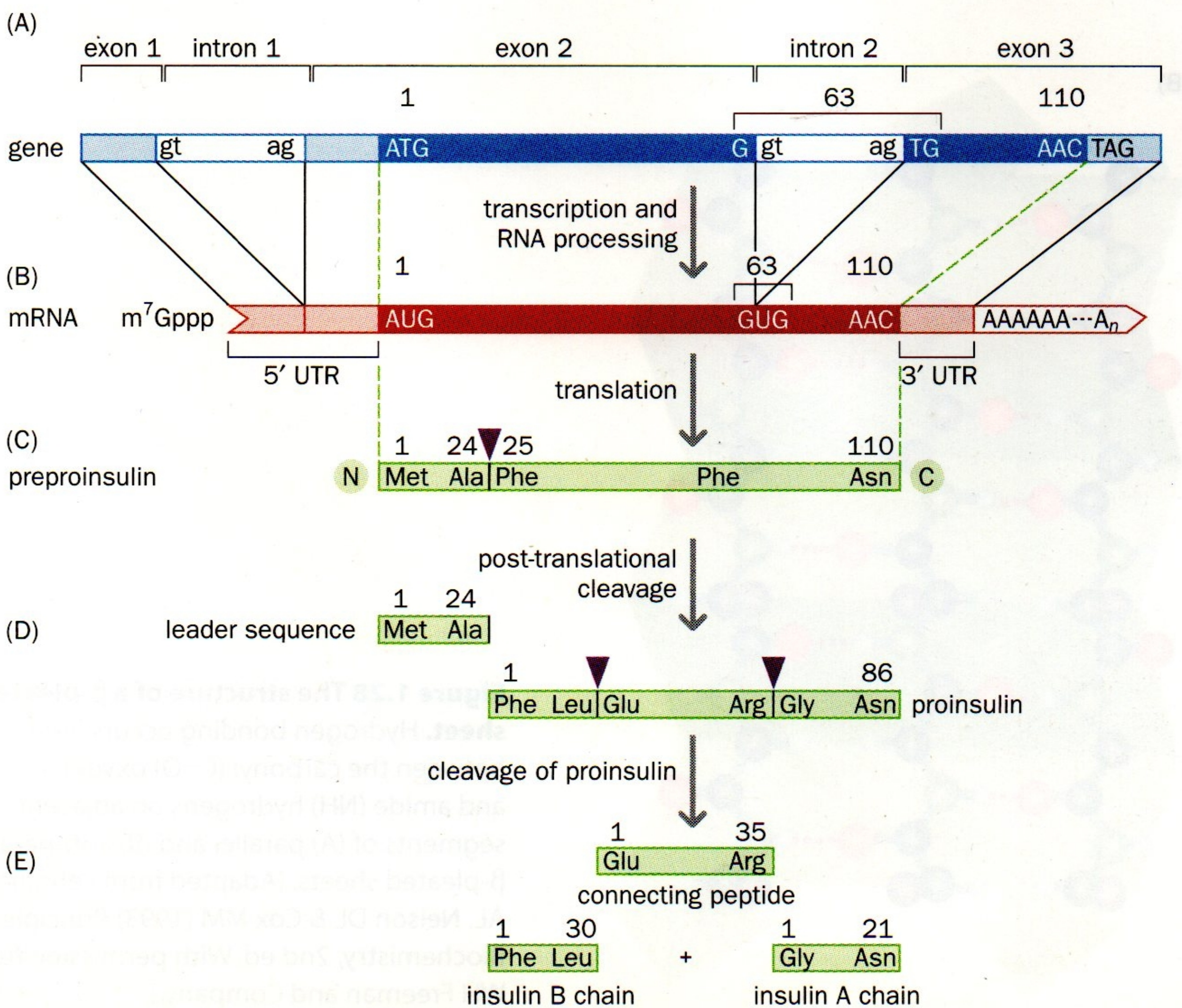


Figure 1.26 Insulin synthesis involves multiple post-translational cleavages of polypeptide precursors. (A) The human insulin gene comprises three exons and two introns. The coding sequence (the part that will be used to make polypeptide) is shown in deep blue. It is confined to the 3′ sequence of exon 2 and the 5′ sequence of exon 3. (B) Exon 1 and the 5′ part of exon 2 specify the 5′ untranslated region (5′ UTR), and the 3′ end of exon 3 specifies the 3′ UTR. The UTRs are transcribed and so are present at the ends of the mRNA. (C) A primary translation product, preproinsulin, has 110 residues and is cleaved to give (D) a 24-residue N-terminal *leader sequence* (that is required for the protein to cross the cell membrane but is thereafter discarded) plus an 86-residue proinsulin precursor. (E) Proinsulin is cleaved to give a central segment (the connecting peptide) that may maintain the conformation of the A and B chains of insulin before the formation of their interconnecting covalent disulfide bridges (see Figure 1.29).

Examples of post-translational processing

Strachan, Read – *Human Molecular Genetics*

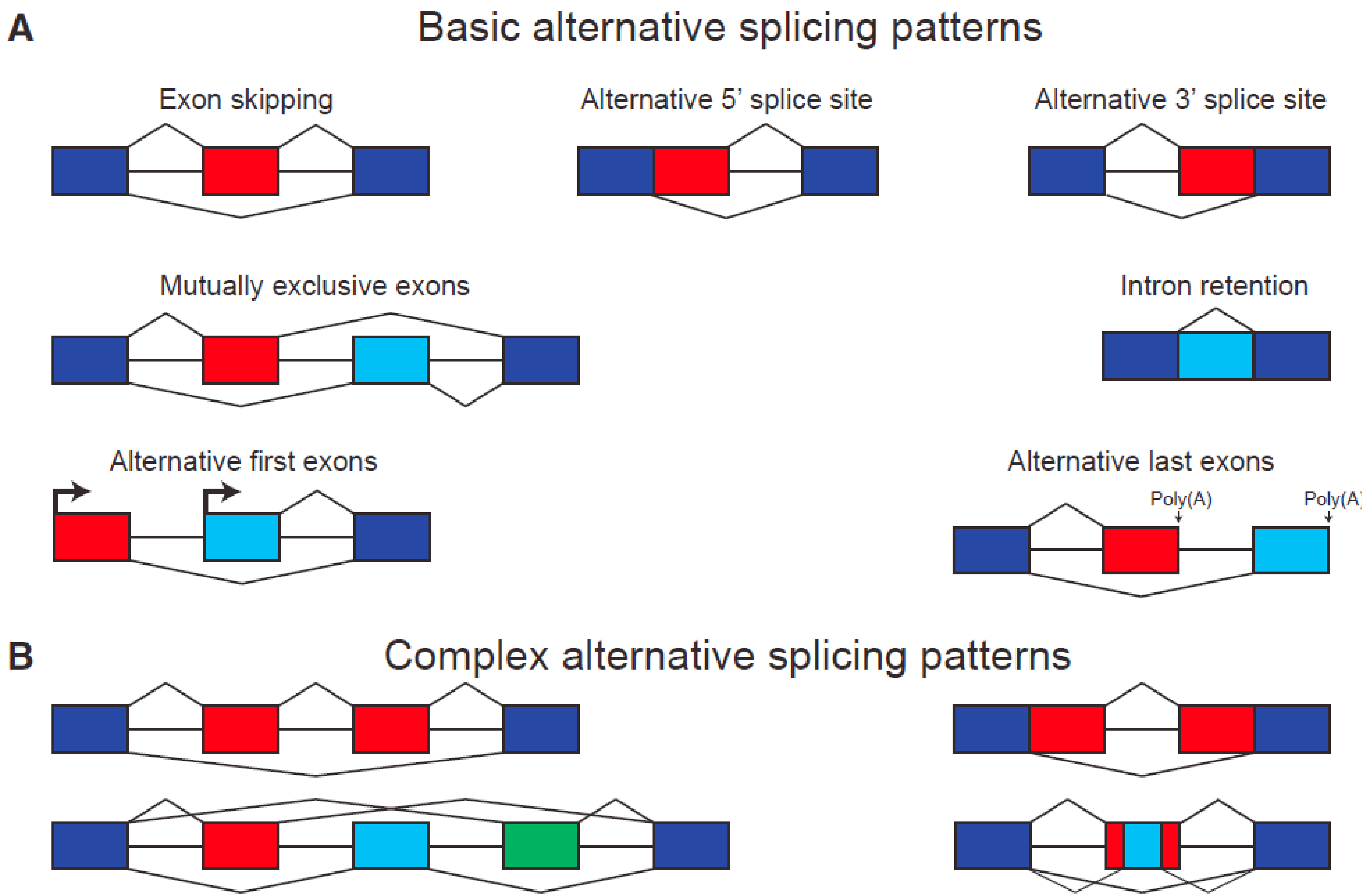
Human gene structure and processing

TABLE 9–1 SOME VITAL STATISTICS FOR THE HUMAN GENOME	
DNA length	3.2×10^9 nucleotide pairs*
Number of genes	approximately 25,000
Largest gene	2.4×10^6 nucleotide pairs
Mean gene size	27,000 nucleotide pairs
Smallest number of exons per gene	1
Largest number of exons per gene	178
Mean number of exons per gene	10.4
Largest exon size	17,106 nucleotide pairs
Mean exon size	145 nucleotide pairs
Number of pseudogenes**	more than 20,000
Percentage of DNA sequence in exons (protein coding sequences)	1.5%
Percentage of DNA in other highly conserved sequences***	3.5%
Percentage of DNA in high-copy repetitive elements	approximately 50%

Q: what gene (exon) is the largest?

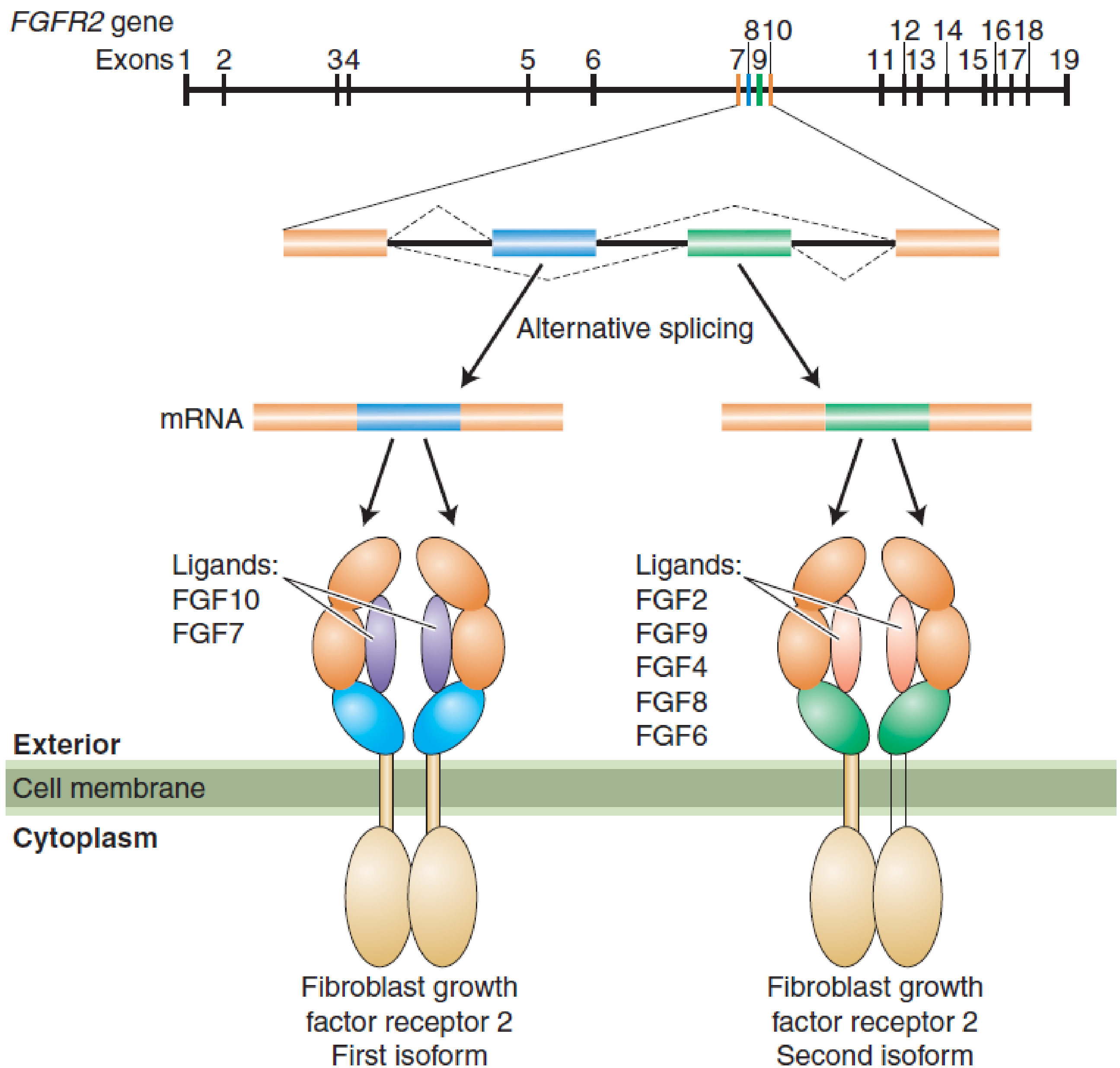
Alberts – *Essential Cell Biology*

Alternative splicing of human genes



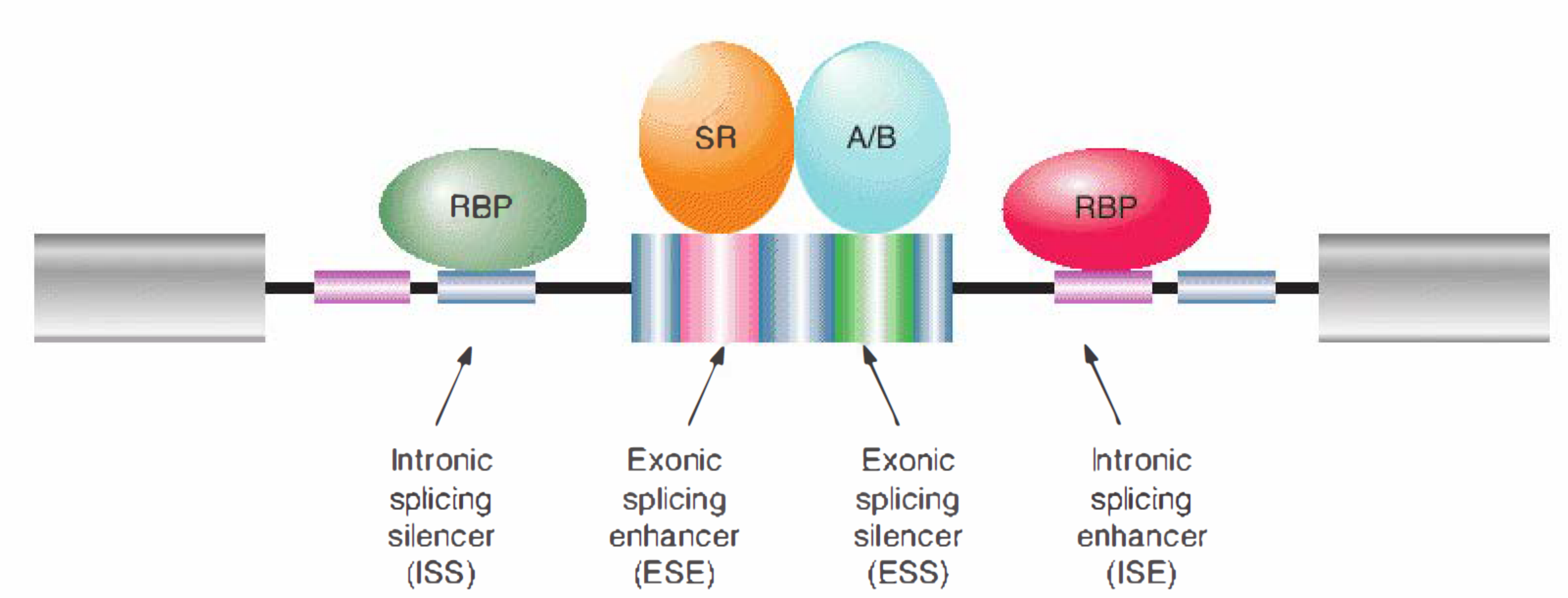
Park (2018) *Am J Hum Genet*

Alternative splicing of human genes

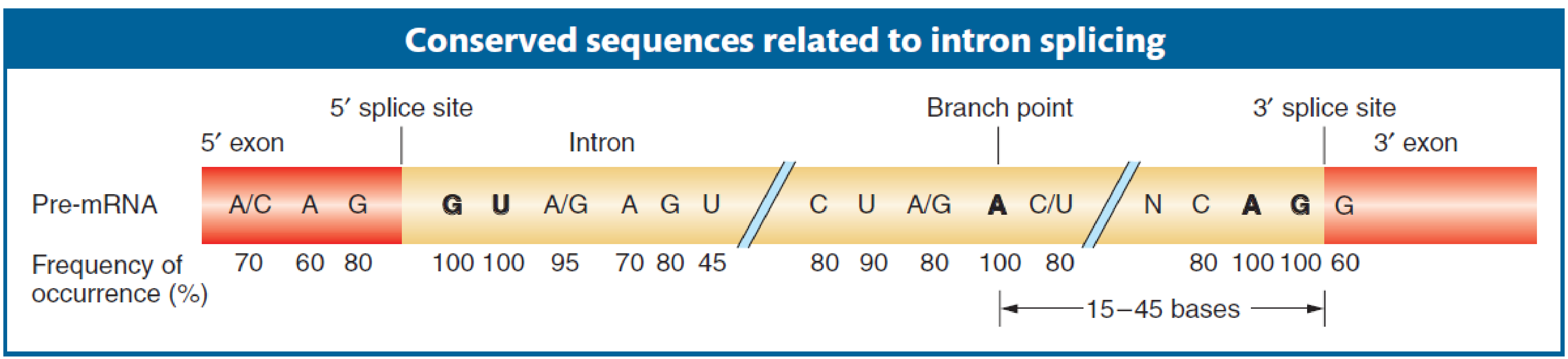


Griffiths -- *Introduction to Genetic Analysis*

Alternative splicing of human genes



Lewin – *Genes XI*



Griffiths -- *Introduction to Genetic Analysis*

Alternative splicing of human genes

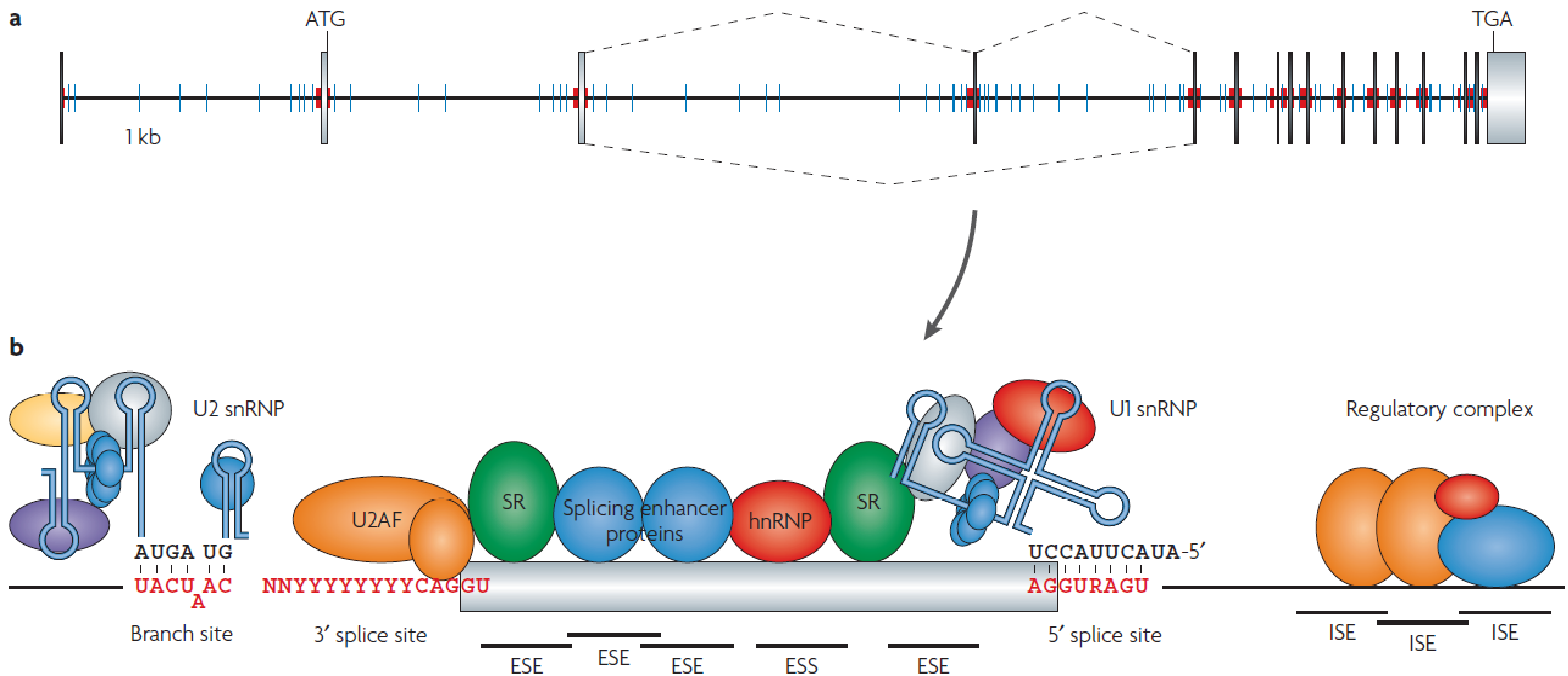


Figure 1 | **The splicing code. a** | A pre-mRNA as it might appear to the spliceosome. Red indicates consensus splice site sequences at the intron–exon boundaries. Blue indicates additional intronic cis-acting elements that make up the splicing code. **b** | cis-elements within and around an alternative exon are required for its recognition and regulation. The 5′ splice site and branch site serve as binding sites for the RNA components of U1 and U2 small nuclear ribonucleoprotein (snRNPs), respectively. This RNA:RNA base pairing determines the precise joining of exons at the correct nucleotides. Mutations in the pre-mRNA that disrupt this base pairing decrease the efficiency of exon recognition. Exons and introns contain diverse sets of enhancer and suppressor elements that refine bone fide exon recognition. Some exon splicing enhancers (ESEs) bind SR proteins and recruit and stabilize binding of spliceosome components such as U2AF. Exon splicing suppressors (ESSs) bind protein components of heterogeneous nuclear ribonucleoproteins (hnRNP) to repress exon usage. Some intronic splicing enhancers (ISEs) bind auxiliary splicing factors that are not normally associated with the spliceosome to regulate alternative splicing.

Wang (2007) *Nat Rev Genet*

Alternative splicing of human genes

- ENSEMBL GRCh38 v.99, protein-coding genes and transcripts:
- 1 transcript: 22.2% (no alternative splicing)
- 2-5 transcripts: 52.9%
- >5 transcripts: 24.9%
- More than 75 transcripts: *ADGRG1, ANK2, KCNMA1, MAPK10, NDRG2, PAX6, TCF4*

- Longest transcript designated as **canonical** (≠ most biologically relevant)
- AS contribution to proteome complexity and transcript functionality is still debated: transcripts ≠ protein isoforms
- AS transcripts that introduce premature stop codon are subject to NMD (**nonsense-mediated decay**)
- Microexons (3-30 nt): misregulated in autistic brain (Irimia (2014) *Cell*).

Aberrant splicing in disease

- **Cis-acting splicing mutations**: disruption of the splicing code, **15-60% of human disease mutations** (Wang 2007 *Nat Rev Genet*)

Examples: synonymous mutations in *CFTR* ⟹ cystic fibrosis;

Splice site mutations in *MITF* ⟹ Waardenburg syndrome type 2 (WS2), a dominantly inherited syndrome of hearing loss and pigmentary disturbances

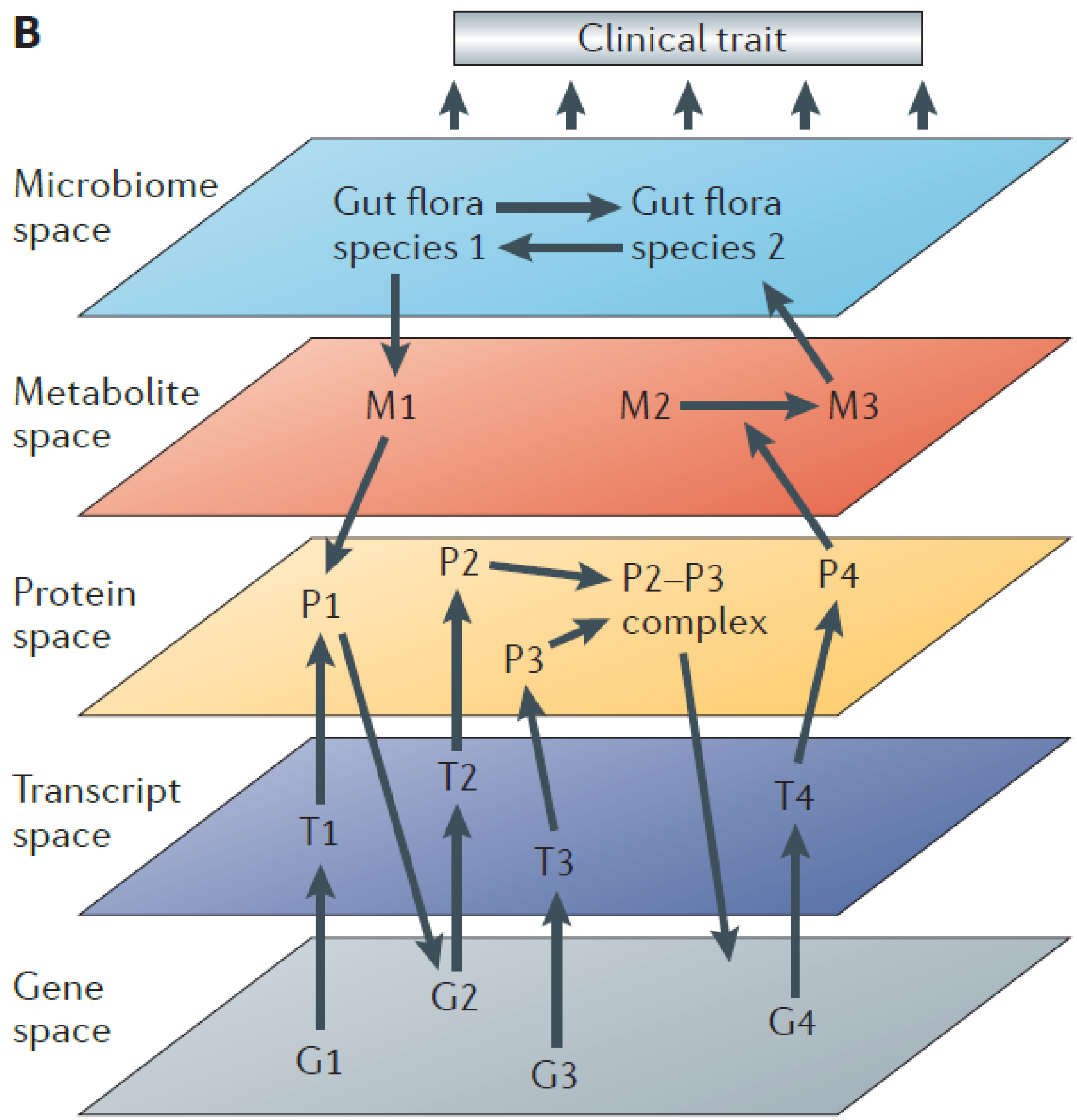
- **Trans-acting mutations**: disruption of the splicing RNA-protein machinery.

Example: mutations in *SMN1* ⟹ loss of snRNP production ⟹ spinal muscular atrophy (SMA). Nusinersen, an antisense oligonucleotide drug for correcting splicing in spinal muscular atrophy.

Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am. J. Hum. Genet*. 102, 11–26.

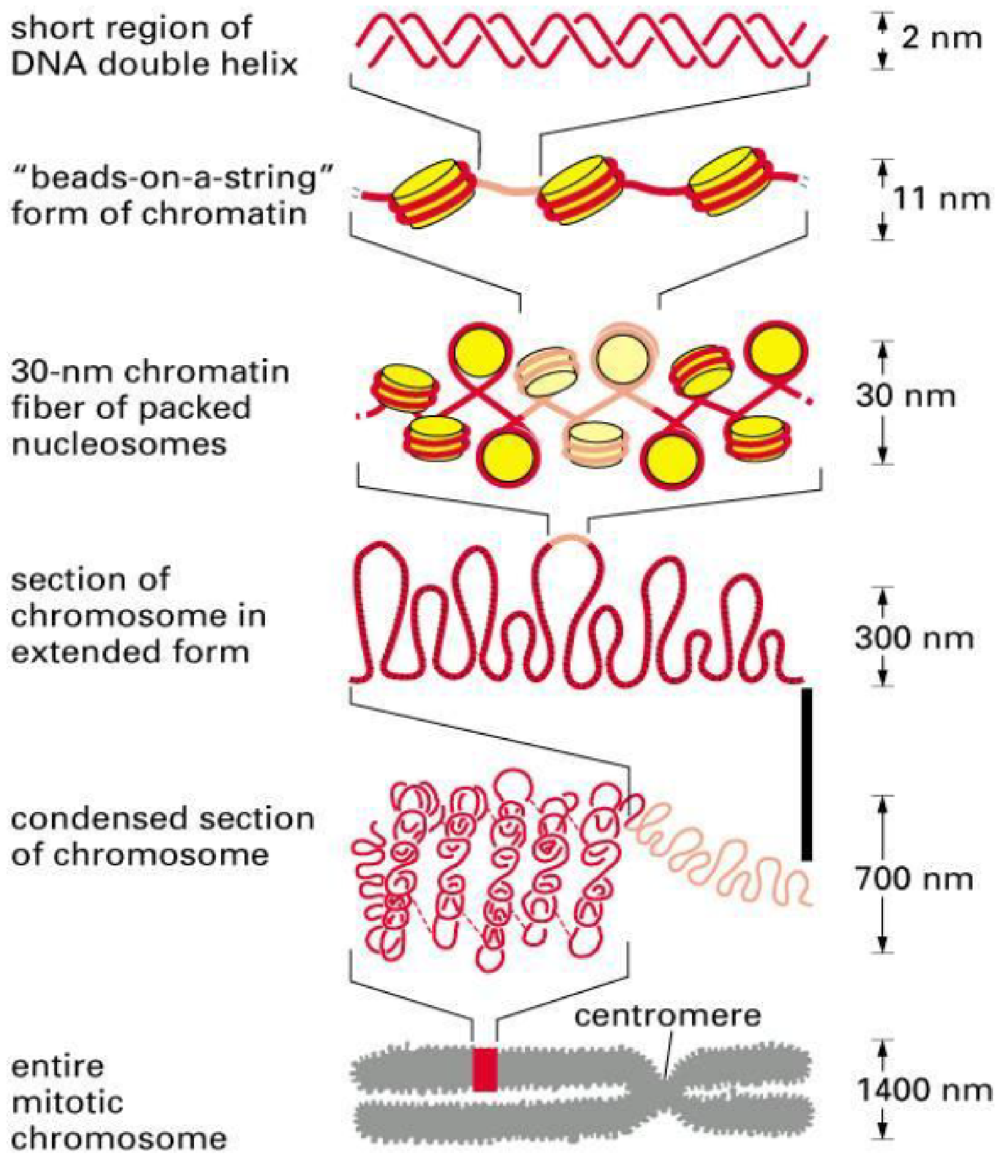
Wang, G.-S., and Cooper, T.A. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. *Nat. Rev. Genet*. 8, 749–761.

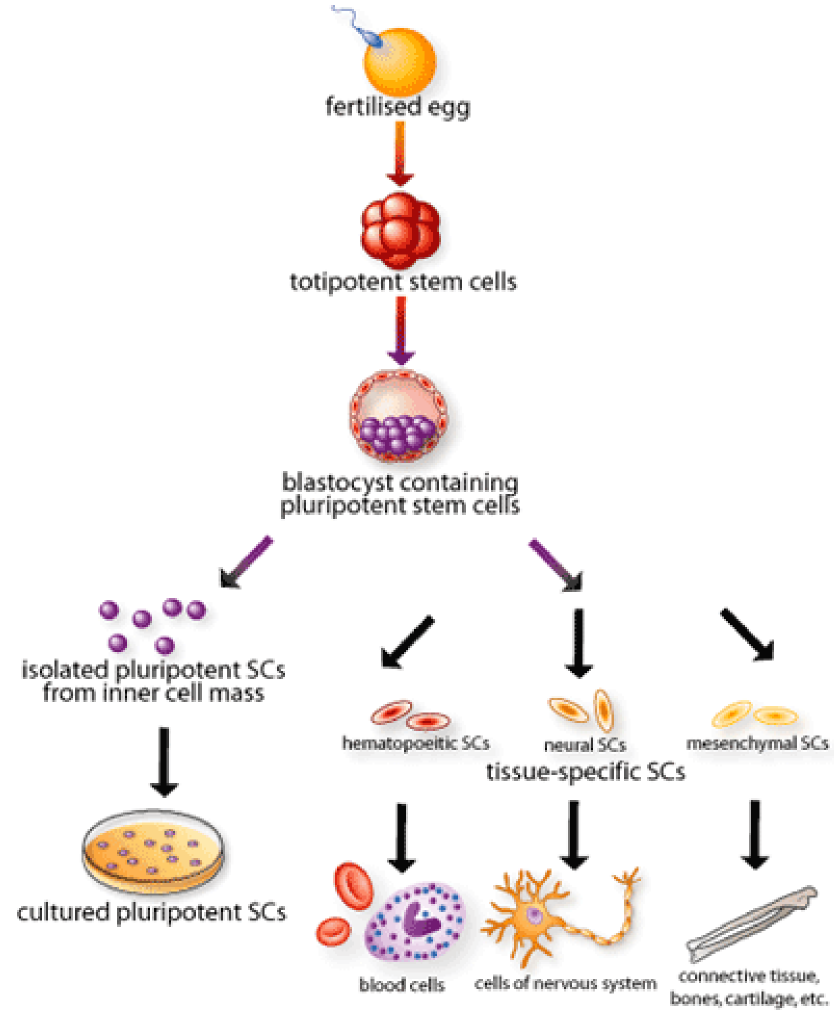
Human genome in action



Civelek (2014) *Nat Rev Genet*

More realistic picture





Molecular Biology of the Cell, 4th ed.

Chaudrey (2004) *Stem Cell Bioeng*

Epigenetics

Epigenetics: heritable phenotype changes that do not involve alterations in the DNA sequence

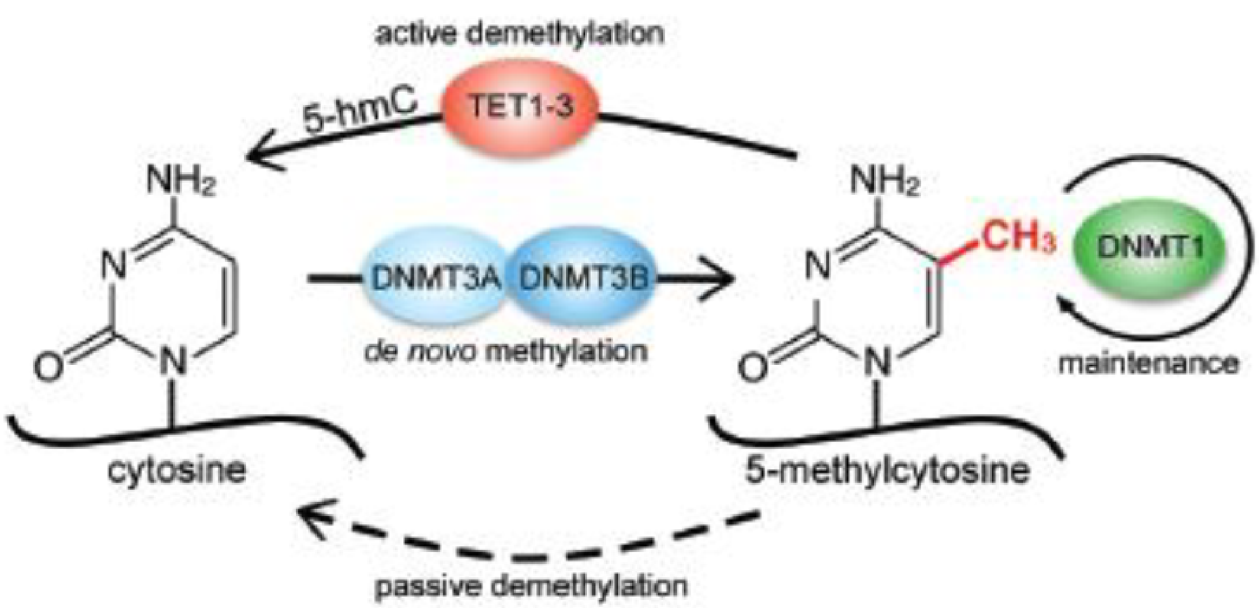
Epigenetic regulation:

1. DNA methylation at CpG dinucleotides
2. Covalent modification of histone proteins
3. Noncoding RNAs

- *Above the genetis*: instructions on using instructions, or gene expression control mechanisms
- Methylation and histone modifications are reversible
- Maintained at cell division and erased during early embriogenesis
- Affected by internal (development, aging) and environmental (chemicals, drugs, diet, lifestyle) factors

DNA methylation

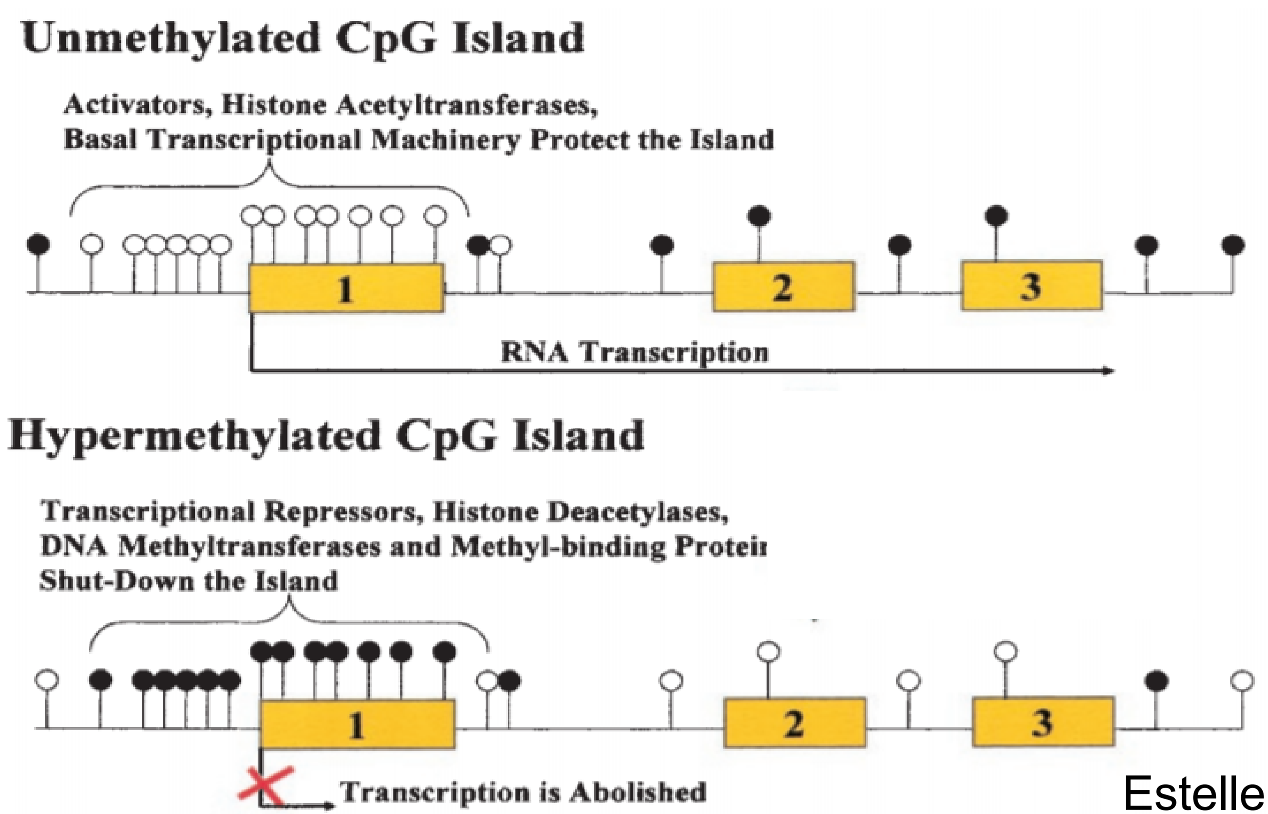
- The only known epigenetic modification of DNA in mammals is methylation of cytosine at position C_5 in CpG dinucleotides
- DNA methyltransferases (DNMTs) establish and maintain DNA methylation patterns
- Methyl-CpG binding proteins (MBDs) read them
- Patterns of CpG methylation may be person-specific, tissue-specific, or locus-specific



Ambrosi (2017) *J Mol Biol*

CpG dinucleotides and islands

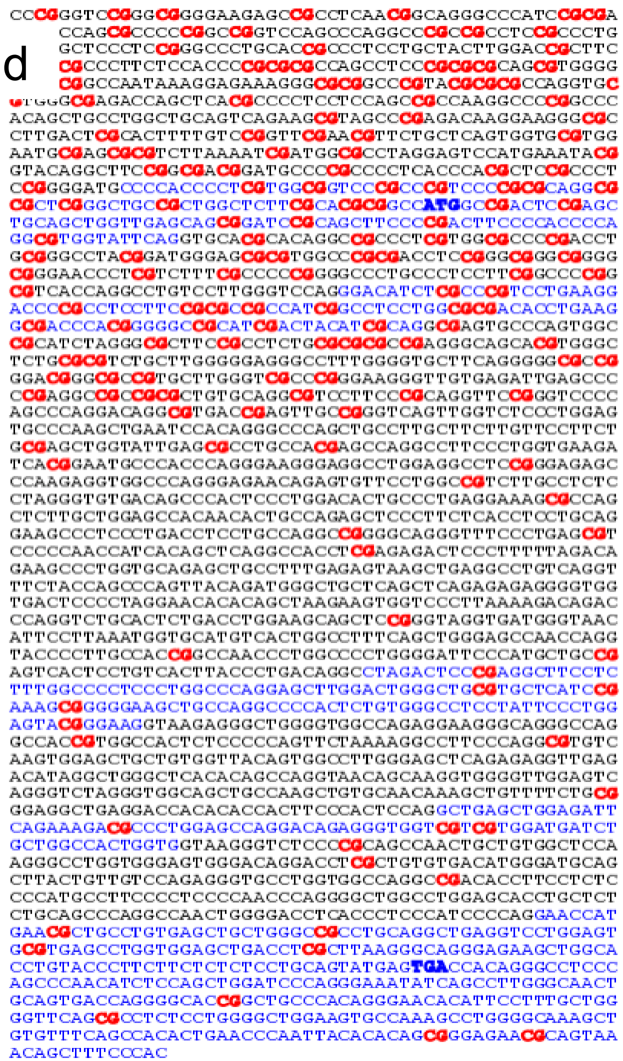
- **CpG island** *ad hoc* definition: length >200 bp, CG >50%, observed-to-expected CpG ratio >60%
- ~30,000 CpG islands in the human genome
- ~70% of human promoters have high CpG content (Saxonov 2006 *PNAS*)
- **Methylation of CpG islands silences gene expression**



Esteller (2002) *Oncogene*

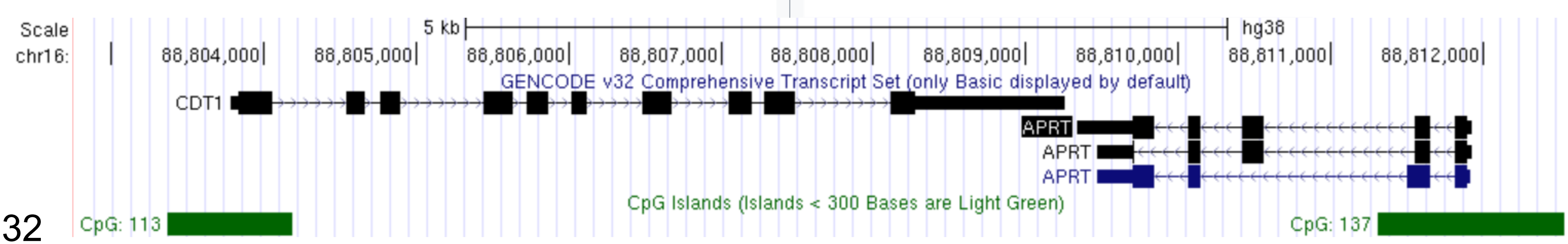
CpG dinucleotides and islands

Rev strand

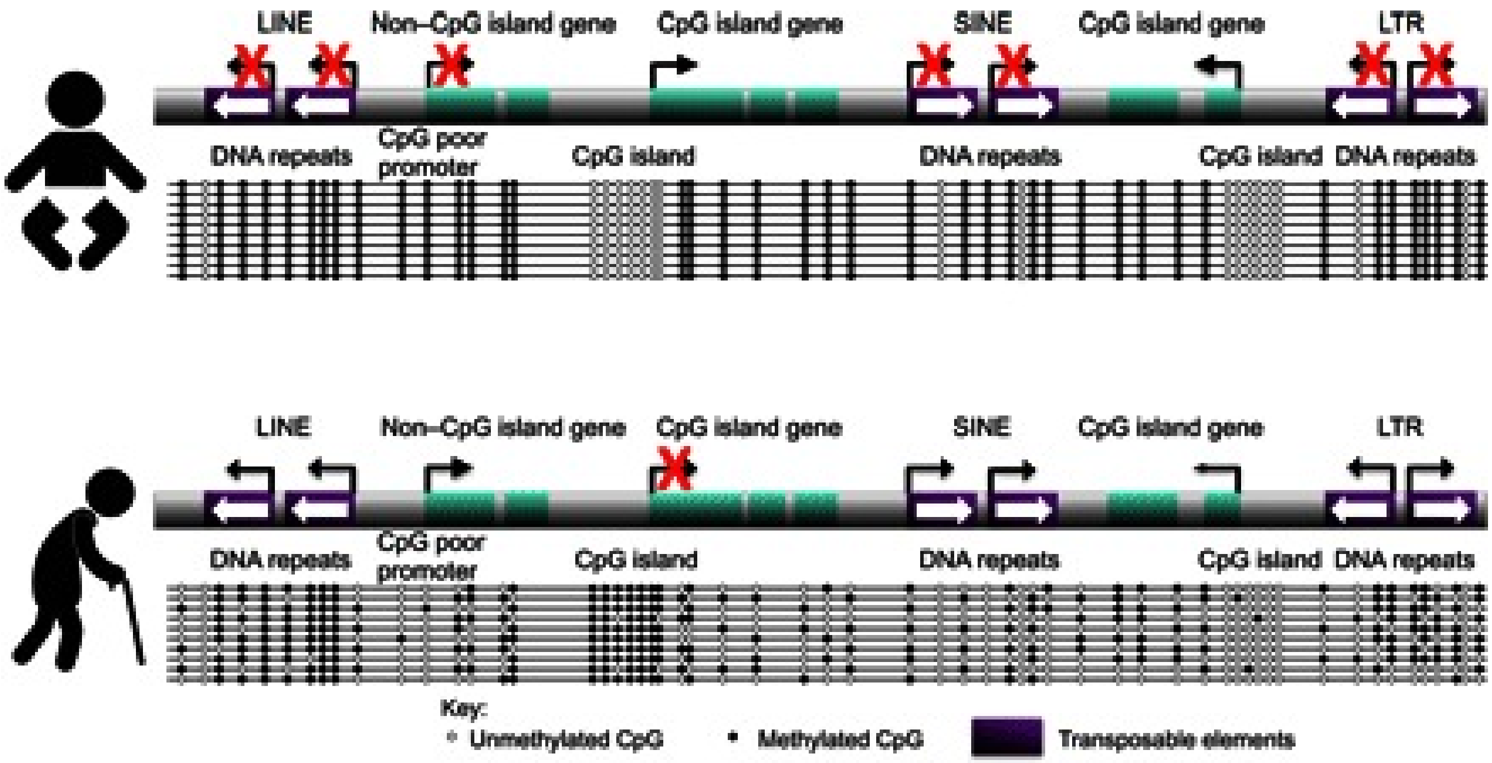


Fwd strand

APRT: Adenine Phosphoribosyltransferase



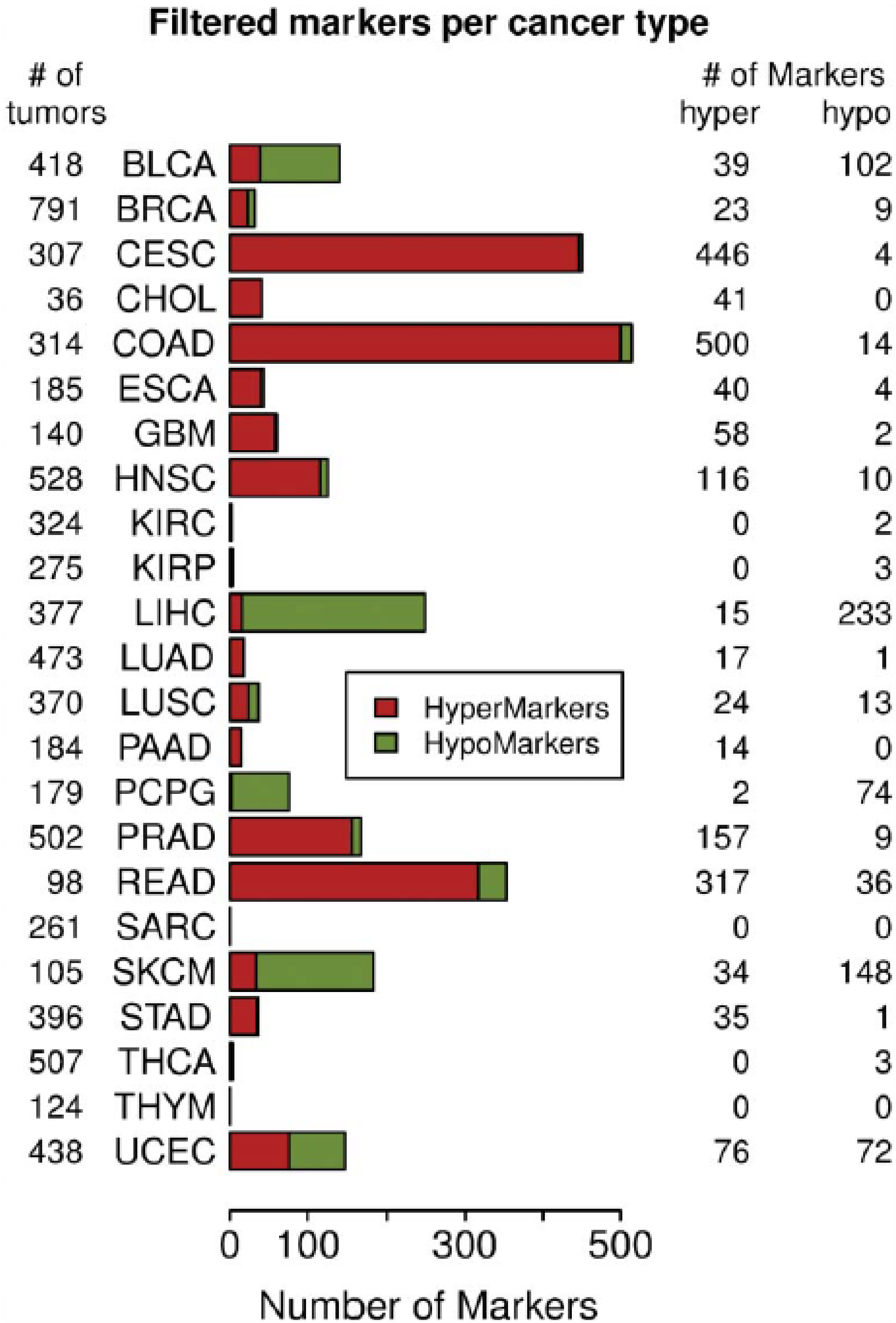
DNA methylation and aging



Young mammalian cells are characterized by DNA hypermethylation over the genome, with the exception of CpG islands within the promoters of expressed genes. In particular, DNA repeats, such as LINE, SINE, and long terminal repeat (LTR) transposable elements, are heavily DNA-methylated, helping to maintain them in a constitutive heterochromatin state. **During aging, there is general DNA hypomethylation over the genome, which mostly occurs in a stochastic manner within the cell population.** Loss of DNA methylation leads to activation of normally silenced DNA sequences like the transposable elements. However, DNA methylation also increases in a nonstochastic manner over the CpG islands of certain genes, correlating with their heterochromatinization and silencing.

Pal & Tyler (2016) *Sci Adv*

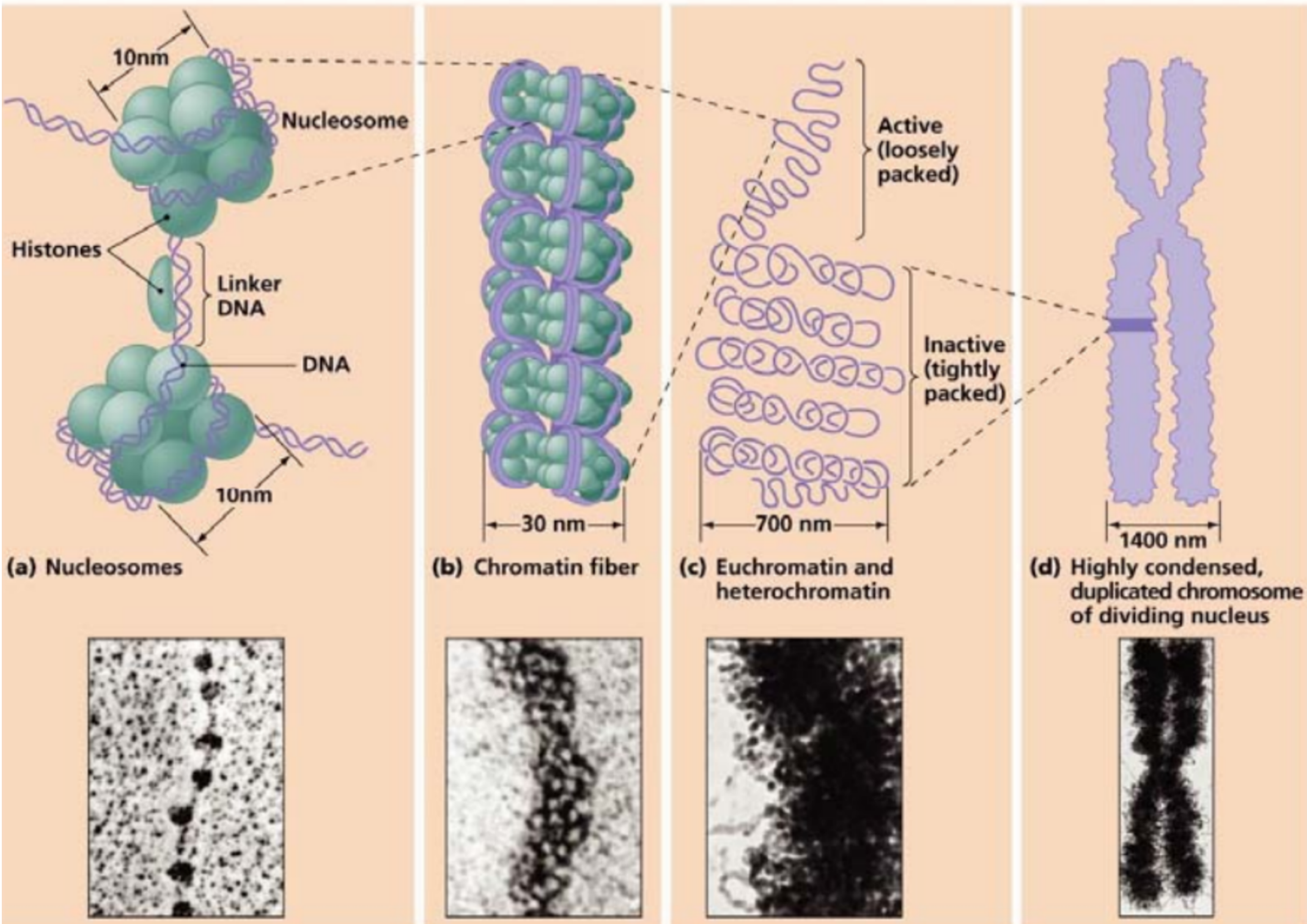
DNA methylation and cancer



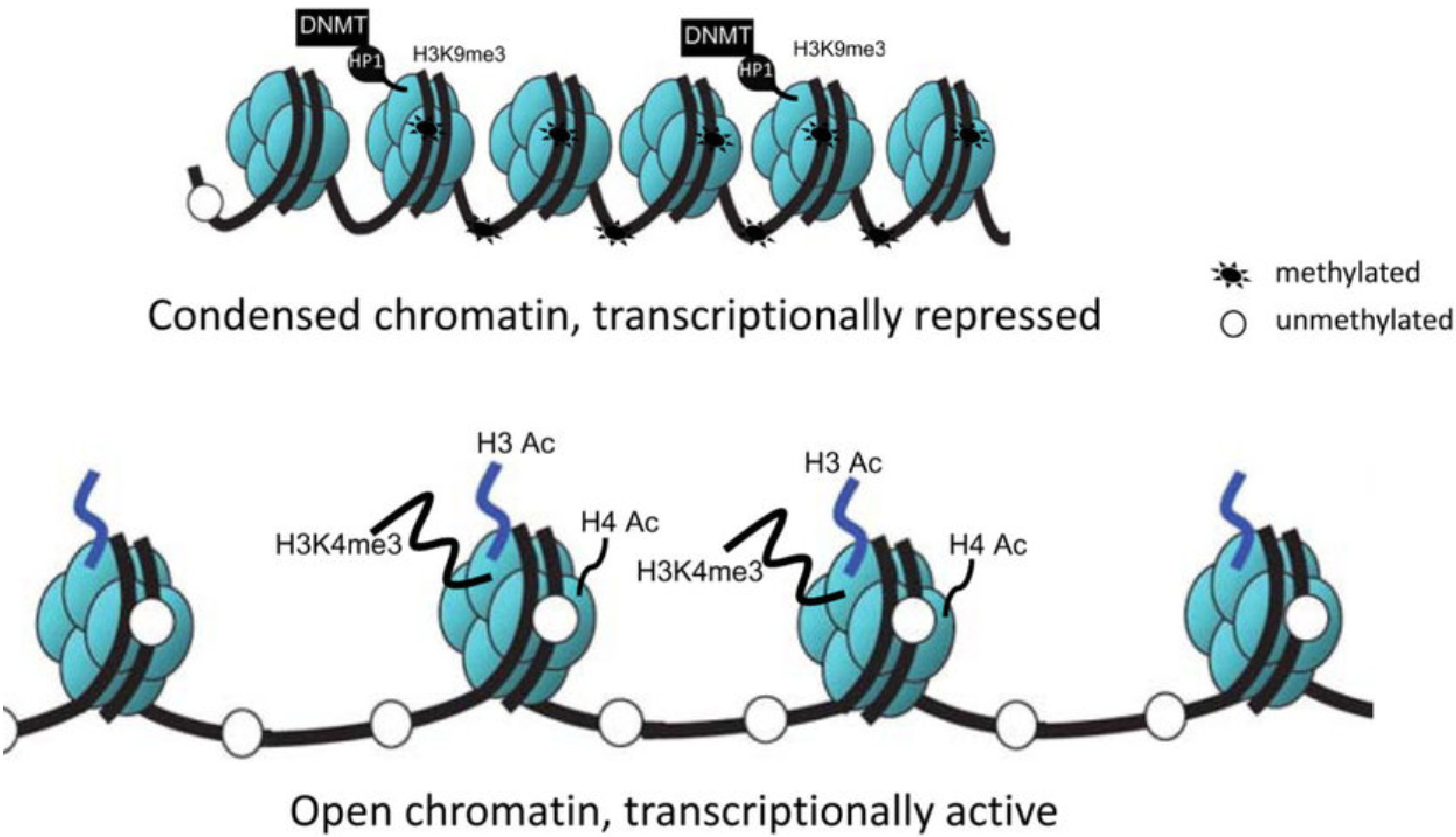
We identified **differentially methylated regions for individual cancer types** and those were further filtered against data from normal tissues to obtain marker regions with cancer-specific methylation, resulting in a total of 1,250 hypermethylated and 584 hypomethylated marker CpGs. From hypermethylated markers, optimal sets of six markers for each TCGA cancer type were chosen that could identify most tumors with high specificity and sensitivity [area under the curve (AUC): 0.969-1.000] and a universal 12 marker set that can detect tumors of all 33 TCGA cancer types (AUC >0.84).

Vrba & Futscher (2018) *Epigenetics*

Histone modifications, histone code



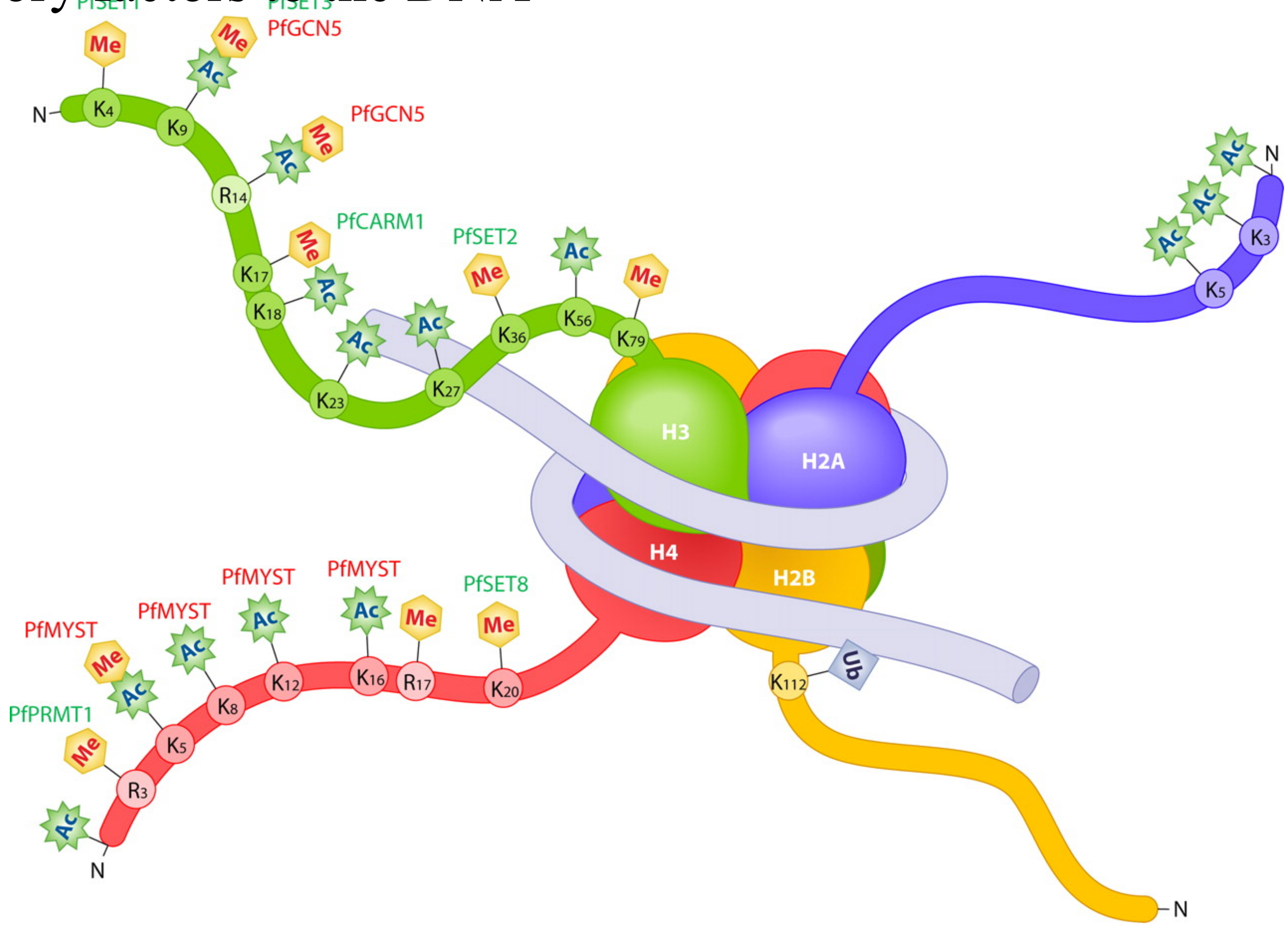
Histone modifications, histone code



Bansal (2017) *Pediatric Diabetes*

Histone modifications, histone code

- **Histone code**: post-translational modifications of histone N-ends (Lys, Arg, Cys) by phosphorylation, acetylation, methylation and ubiquitylation.
- These changes regulate gene expression by modulating the access of regulatory factors to the DNA



themysteryofepigenetics.blogspot.com

Histone modifications, histone code

The eukaryotic genome is organized in what is known as a **nucleosome**, the first level of condensation. The nucleosome is composed of 147 base pairs of negatively-charged DNA wrapped twice around an octamer of positively-charged proteins called **histones**. It consists of two H2A and H2B dimers, and a H3 and H4 tetramer. The nucleosomes are separated by 1,016 base pairs (bp) of DNA called "linker DNA", which constitutes an arrangement referred to as "beads on a string", that is around 10nm in diameter. DNA can be further condensed at different points during the cell cycle, forming a 30nm chromatin fiber composed of packed nucleosomes using the histone H1, which binds to the linker DNA. These 30nm fibers can form scaffolds and further condense until chromosomes are formed, which are the highest form of DNA organization within a cell.

Histones have very dynamic N-terminal "tails" extending from the surface of the nucleosome that are rich in basic amino acids. These tails can be modified by post-translational modifications (PTM's) catalyzed by a variety of enzymes, by adding either methyl, acetyl or phosphoryl groups. Aditionally, lysines can be mono, di or tri-methylated, while arginine can accept up to two methyl groups which adds to the complexity. Methylation of DNA at cytosine residues, as well as PTMs of histones, including phosphorylation, acetylation, methylation and ubiquitylation, contributes to the epigenetic information carried by chromatin. These changes play an important role in the regulation of gene expression by modulating the access of regulatory factors to the DNA. Many modification sites are close enough to each other and it seems that modification of histone tails by one enzyme might influence the rate and efficiency at which other enzymes use the newly modified tails as a substrate.

Histone modifications, histone code

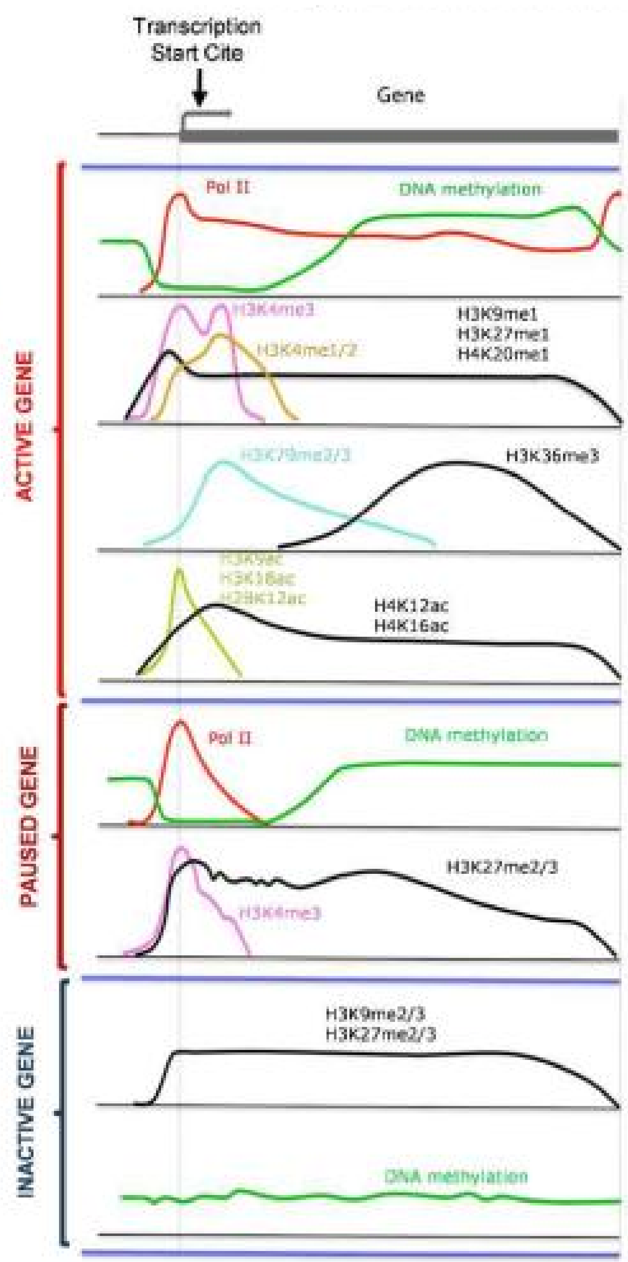
Table 1. The histone code.

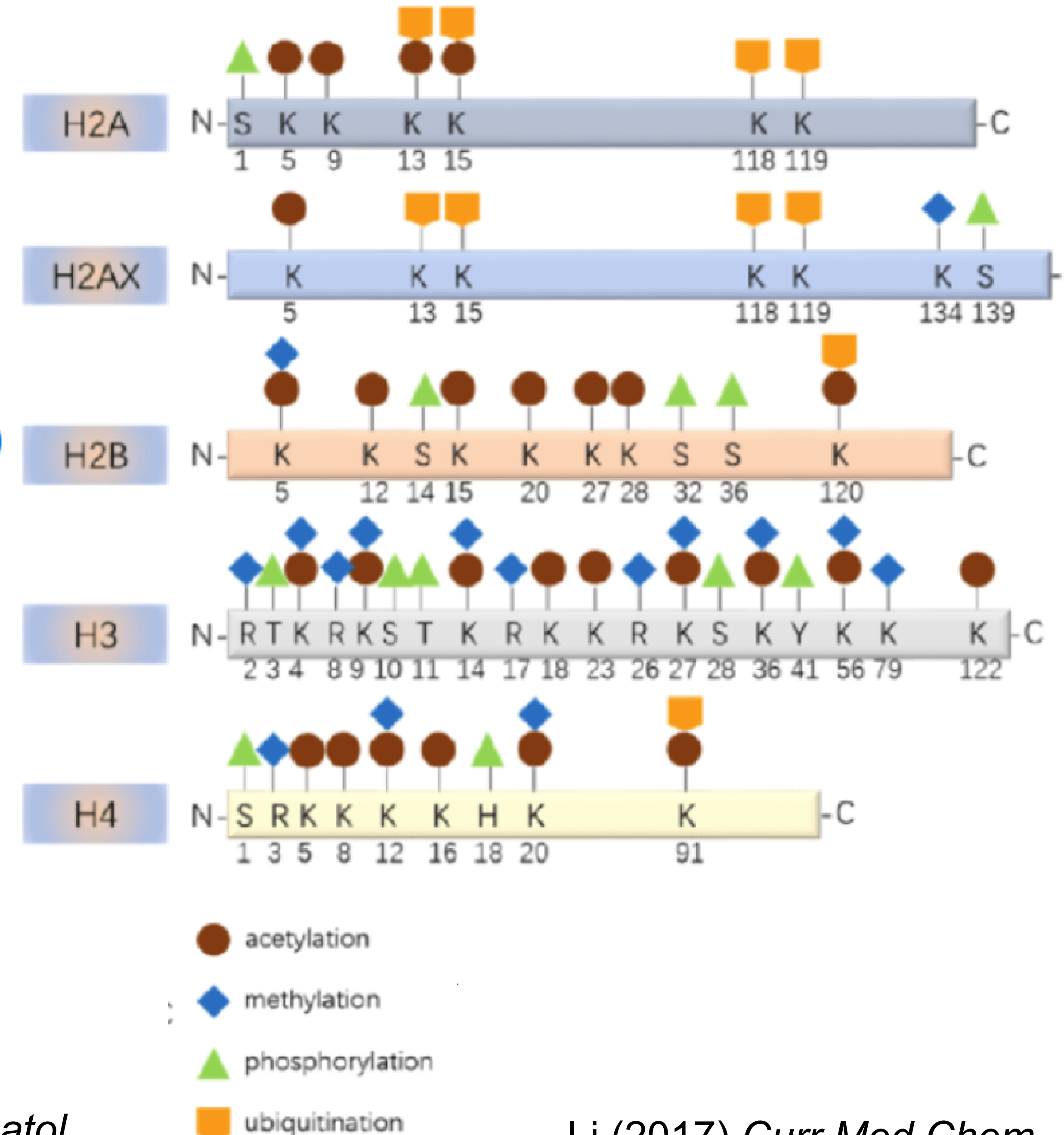
Histone code	Methylation			Acetylation	Ubiquitination
	Monomethylation	Dimethylation	Trimethylation		
H2AK119	–	–	–	–	Repression
H2BK5	Activation	–	Repression	–	–
H3K4	Activation	Activation	Activation	–	–
H3K9	Activation	Repression	Repression	Activation	–
H3K14	–	–	–	Activation	–
H3K18	–	–	–	Activation	–
H3K27	Activation	Repression	Repression	Activation	–
H3K36	Repression	Activation	Activation	–	–
H3K56	–	–	–	Activation	–
H3K79	Activation	Activation	Activation, repression	–	–
H4K12	–	–	–	Activation	–
H4K20	Activation		Repression	–	–

For each post-translational modification, the known functional association on gene transcription is shown. By reading the combinatorial and/or sequential histone modifications that constitute the histone code, it may be possible to predict which gene products will be transcribed. However, this code is controversial, since some gene loci present marks both associated with transcriptional activation and linked with repression. These bivalent domains are posited to be poised for either up- or down-regulation and to provide an epigenetic blueprint for lineage determination, and are usually found in stem cells.

Bauge (2014) *Future Med Chem*

Histone modifications, histone code



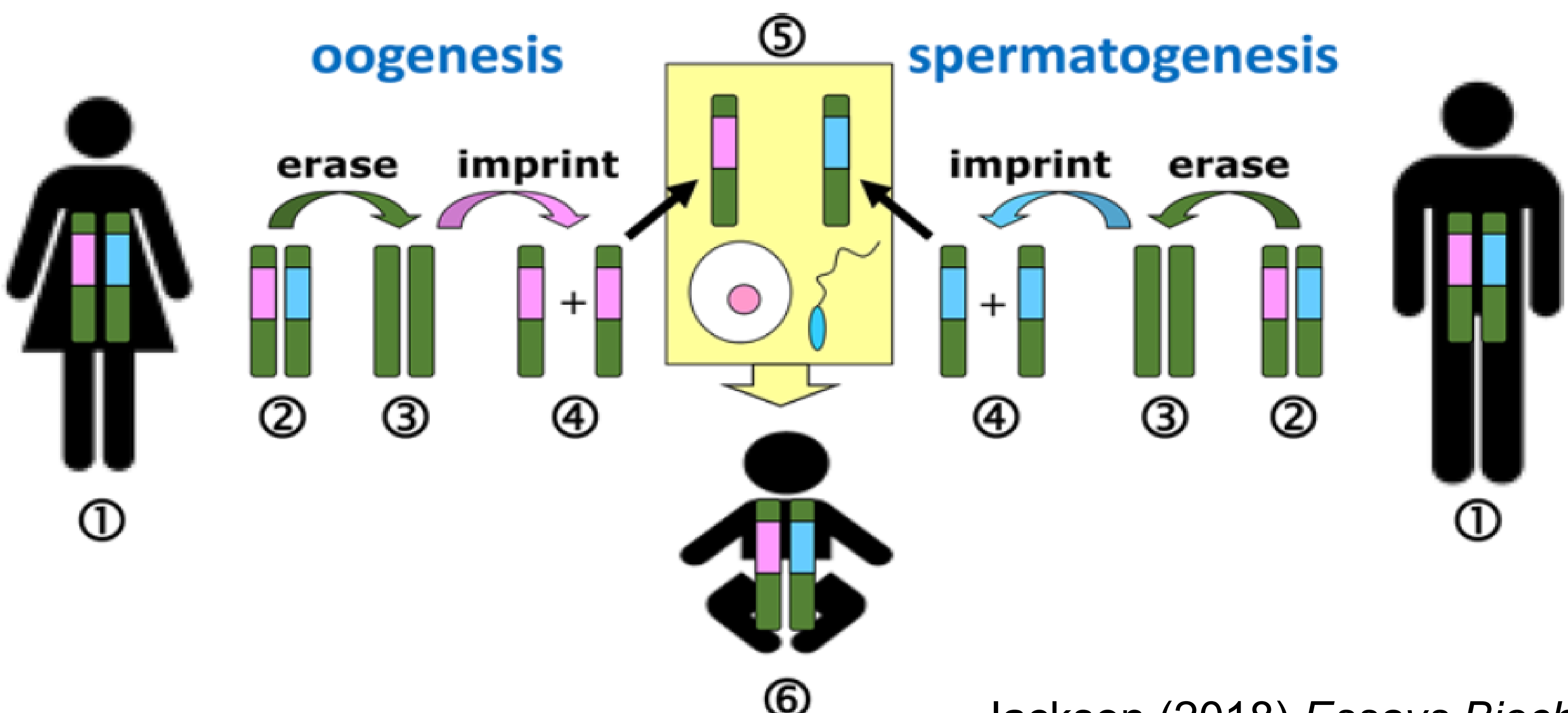


 Botchkarev (2012) *J Invest Dermatol*

Li (2017) *Curr Med Chem*

Chromosomal imprinting

- **Chromosomal imprinting, or imprints**: ~100 genes on various chromosomes, one copy is inactive by epigenetic mechanisms depending upon parent of origin
- For some genes (~70) only the paternal allele is active, while the maternal copy is epigenetically silenced throughout the life of the individual, and vice versa (~30 genes)
- Mutations in an active copy of a gene result in **imprinting disorders**

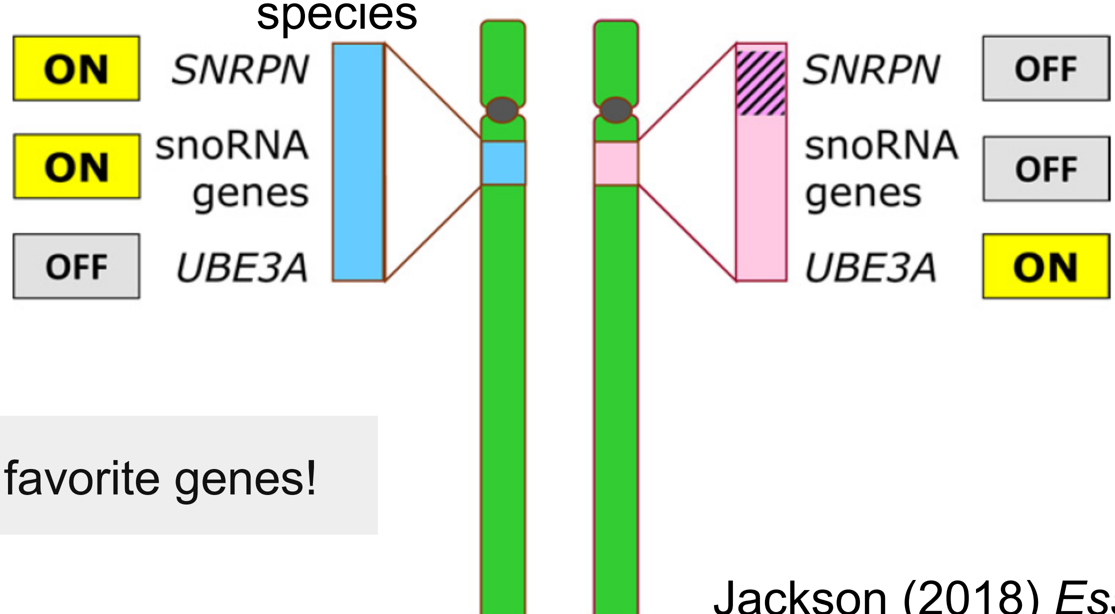


Jackson (2018) *Essays Biochem*

Chromosomal imprinting

Gene	Aliases	Location	Status	Expressed Allele
MAGEL2	nM15, NDNL1	15q11-q12 *AS*	**Imprinted**	Paternal
MKRN3	D15S9, RNF63, ZFP127, ZNF127, MGC88288	15q11-q13	**Imprinted**	Paternal
UBE3A	AS, ANCR, E6-AP, HPVE6A, EPVE6AP, FLJ26981	15q11-q13 *AS*	**Imprinted**	Maternal
NPAP1	C15orf2	15q11-q13	**Imprinted**	Unknown
ZNF127AS	MKRN3AS, Znp127as	15q11-q13	**Unknown**	Unknown
SNORD109A	HBII-438A	15q11.2	**Imprinted**	Paternal
SNORD108	HBII-437, HBII-437 C/D box snoRNA	15q11.2	**Imprinted**	Paternal
SNORD107	HBII-436, HBII-436 C/D box snoRNA	15q11.2	**Imprinted**	Paternal
SNORD109B	HBII-438B, HBII-438B C/D box snoRNA	15q11.2	**Imprinted**	Paternal
ATP10A	ATPVA, ATPVC, ATP10C, KIAA0566	15q11.2 *AS*	**Imprinted**	Maternal
SNRPN	SMN, PWCR, SM-D, RT-LI, HCERN3, SNRNP-N, FLJ33569, FLJ36996, FLJ39265, MGC29886, SNURF-SNRPN, DKFZp762N022, DKFZp686C0927, DKFZp761I1912, DKFZp686M12165	15q11.2	**Imprinted**	Paternal

http://www.geneimprint.com/site/genes-by-species



Exercise: check your favorite genes!

Jackson (2018) *Essays Biochem*

Imprinting disorders

	Angelman syndrome	*Prader-Willi syndrome*
Key features	* Moderate to severe ID (IQ ~25–54) * Jerky, puppet-like movements * Happy and sociable disposition * Seizures	* Mild to moderate ID (IQ ~60–70) * Insatiable appetite leading to morbid obesity * Behaviour problems
Frequency in the population	~1/20,000	~1/15,000
Underlying genetic abnormality (in some cases, the underlying cause has not been determined)	– Maternal 15q11.2 deletion (~70%) – Paternal UPD (~4%) – Imprinting defect (~8%) – Pathogenic variant in UBE3A (~6%)	– Paternal 15q11.2 deletion (~70%) – Maternal UPD (~20%) – Imprinting defect (~5%)
Key genes	*UBE3A* encoding a ubiquitin ligase	*SNORD116* gene cluster encoding snoRNAs (other genes in the imprinted region may also influence the phenotype)

Jackson (2018) *Essays Biochem*

Imprinting disorders

- IGF2 is a hormone that stimulates growth during embryonic and fetal development // not the IGF2 receptor gene!
- Normally maternally silenced in humans
- **Epimutation** (missing methyl tags) can result in two active copies

Activation of the maternal *IGF2* gene during egg formation or very early in development causes **Beckwith-Wiedemann Syndrome (BWS):**

– overgrowth
– an increased risk of cancer, especially during childhood
– variety of other symptoms

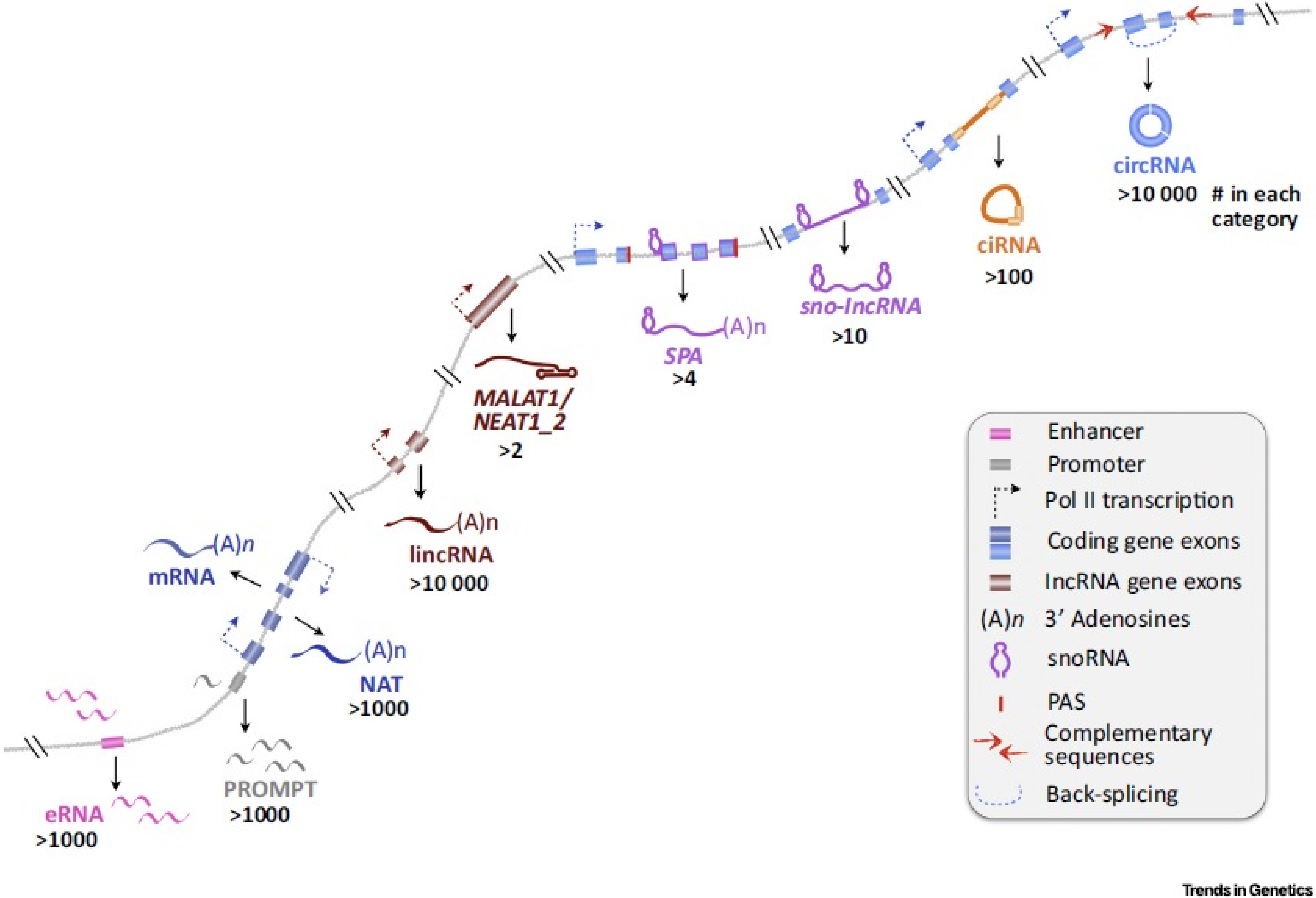
Beckwith-Wiedemann syndrome

Macroglossia Umbilical hernia Omphalocele

Frequency: ~15,000 births. However, in babies that were conceived in the laboratory with the help of artificial reproductive technology, the rate of BWS may be as high as 1/4,000.

https://learn.genetics.utah.edu/content/epigenetics/imprinting

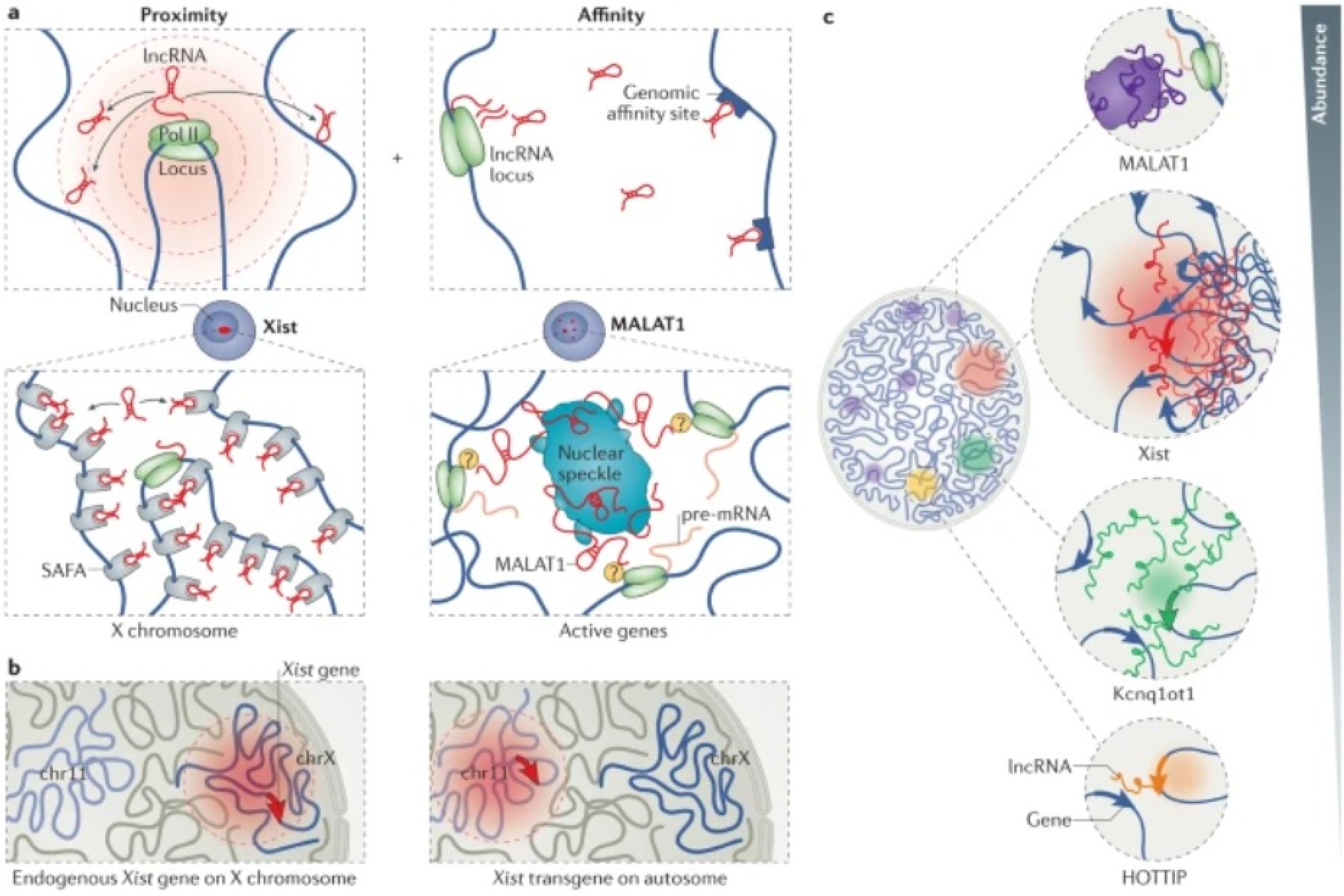
Non-coding RNAs in the genome



Huang Wu (2017) *Trends Genet*

Non-coding RNAs in the genome

Mechanisms of long non-coding RNA localization to chromatin



 Engreitz (2016) *Nat Rev Mol Cell Biol*

Non-coding RNAs in the genome

Name	Size	Location	Number in humans	Functions	Illustrative examples
Short ncRNAs					
miRNAs	19–24 bp	Encoded at widespread locations	>1,424	Targeting of mRNAs and many others	miR-15/16, miR-124a, miR-34b/c, miR-200
piRNAs	26–31bp	Clusters, intragenic	23,439	Transposon repression, DNA methylation	piRNAs targeting *RASGRF1* and LINE1 and IAP elements
tiRNAs	17–18bp	Downstream of TSSs	>5,000	Regulation of transcription?	Associated with the *CAP1* gene
Mid-size ncRNAs					
snoRNAs	60–300 bp	Intronic	>300	rRNA modifications	U50, SNORD
PASRs	22–200 bp	5′ regions of protein-coding genes	>10,000	Unknown	Half of protein-coding genes
TSSa-RNAs	20–90 bp	−250 and +50 bp of TSSs	>10,000	Maintenance of transcription?	Associated with *RNF12* and *CCDC52* genes
PROMPTs	<200 bp	−205 bp and −5 kb of TSSs	Unknown	Activation of transcription?	Associated with *EXT1* and *RBM39* genes
Long ncRNAs					
lincRNAs	>200 bp	Widespread loci	>1,000	Examples include scaffold DNA–chromatin complexes	*HOTAIR*, *HOTTIP*, *lincRNA-p21*
T-UCRs	>200 bp	Widespread loci	>350	Regulation of miRNA and mRNA levels?	uc.283+, uc.338, uc160+
Other lncRNAs	>200 bp	Widespread loci	>3,000	Examples include X-chromosome inactivation, telomere regulation, imprinting	*XIST*, *TSIX*, TERRAs, *p15AS*, *H19*, *HYMAI*

Esteller (2011) *Nat Rev Genet*

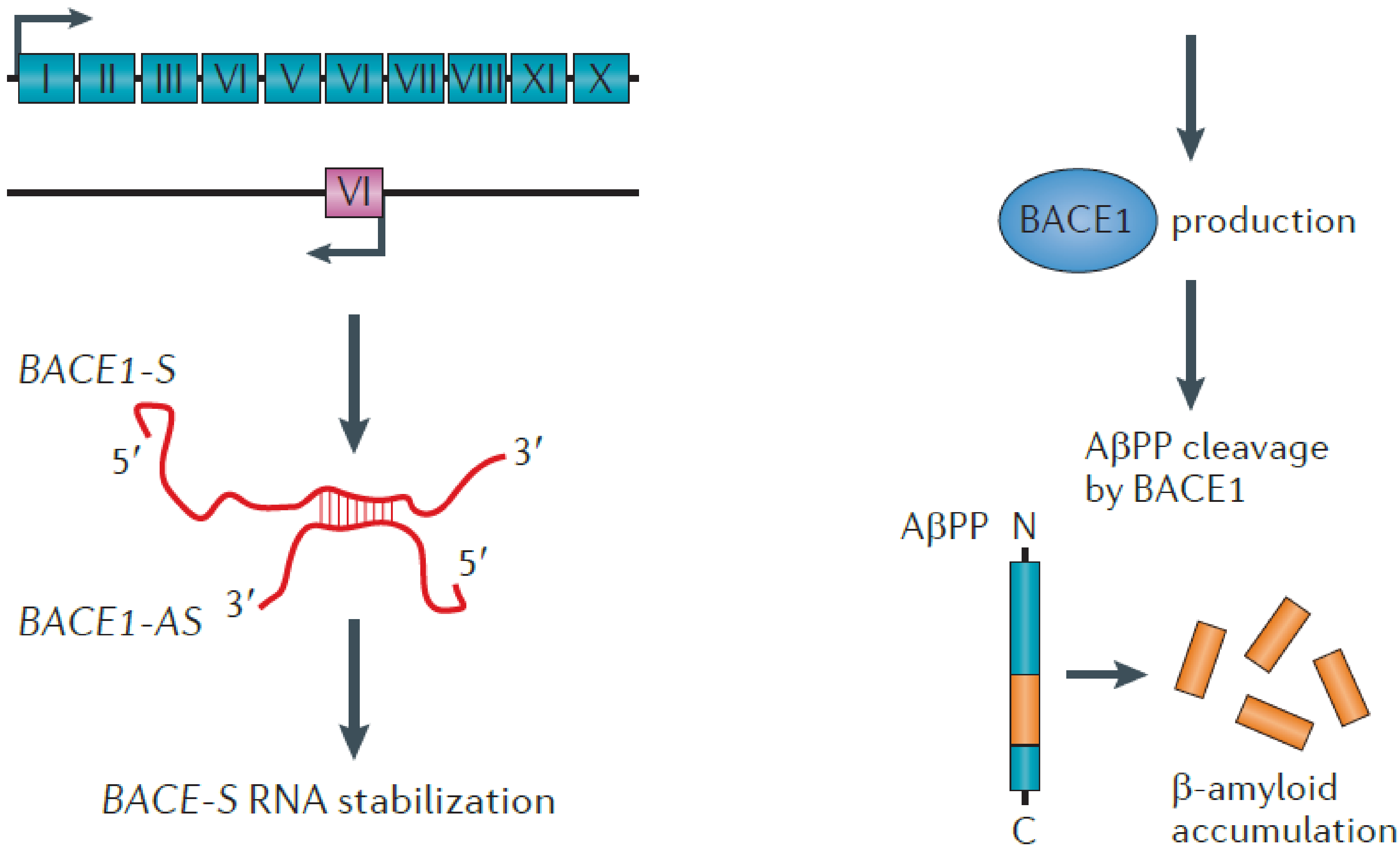
Non-coding RNAs in non-cancer disease

Disease	Involved ncRNAs	ncRNA type
Spinal motor neuron disease	miR-9	miRNA
Spinocerebellar ataxia type 1	miR-19, miR-101, miR-100	miRNA
Amyotropic lateral sclerosis	miR-206	miRNA
Arrhytmia and hypertension	miR-1	miRNA
Atheromatosis and atherosclerosis	miR-10a, miR-145, mR-143 and miR-126	miRNA
Atheromatosis and atherosclerosis	Circular ncRNA linked to the CDKN2A locus	lncRNA
Cardiac hypertrophy	miR-21	miRNA
Rett's syndrome	miR-146a, miR-146b, miR-29 and miR-382	miRNA
5q syndrome	miR-145 and miR-146a	miRNA
ICF syndrome	miR-34b, miR-34c, miR-99b, let-7e and miR-125a	miRNA
Crohn's disease	miR-196	miRNA
Prader–Willi and Angelman syndromes	snoRNA cluster at 15q11–q13 imprinted locus	snoRNA
Beckwith–Wiedeman syndrome	lncRNAs *H19* and *KCNQ1OT1*	lncRNA
Uniparental disomy 14	snoRNA cluster at 14q32.2 imprinted locus	snoRNA
Silver–Russell syndrome	lncRNA *H19*	lncRNA
Silver–Russell syndrome	miR-675	miRNA
McCune–Albright syndrome	lncRNA *NESP-AS*	lncRNA
Deafness	miR-96	miRNA
Alzheimer's disease	miR-29, miR-146 and miR-107	miRNA
Alzheimer's disease	ncRNA antisense transcript for *BACE1*	lncRNA

Exercise: research a ncRNA-related disease

Esteller (2011) *Nat Rev Genet*

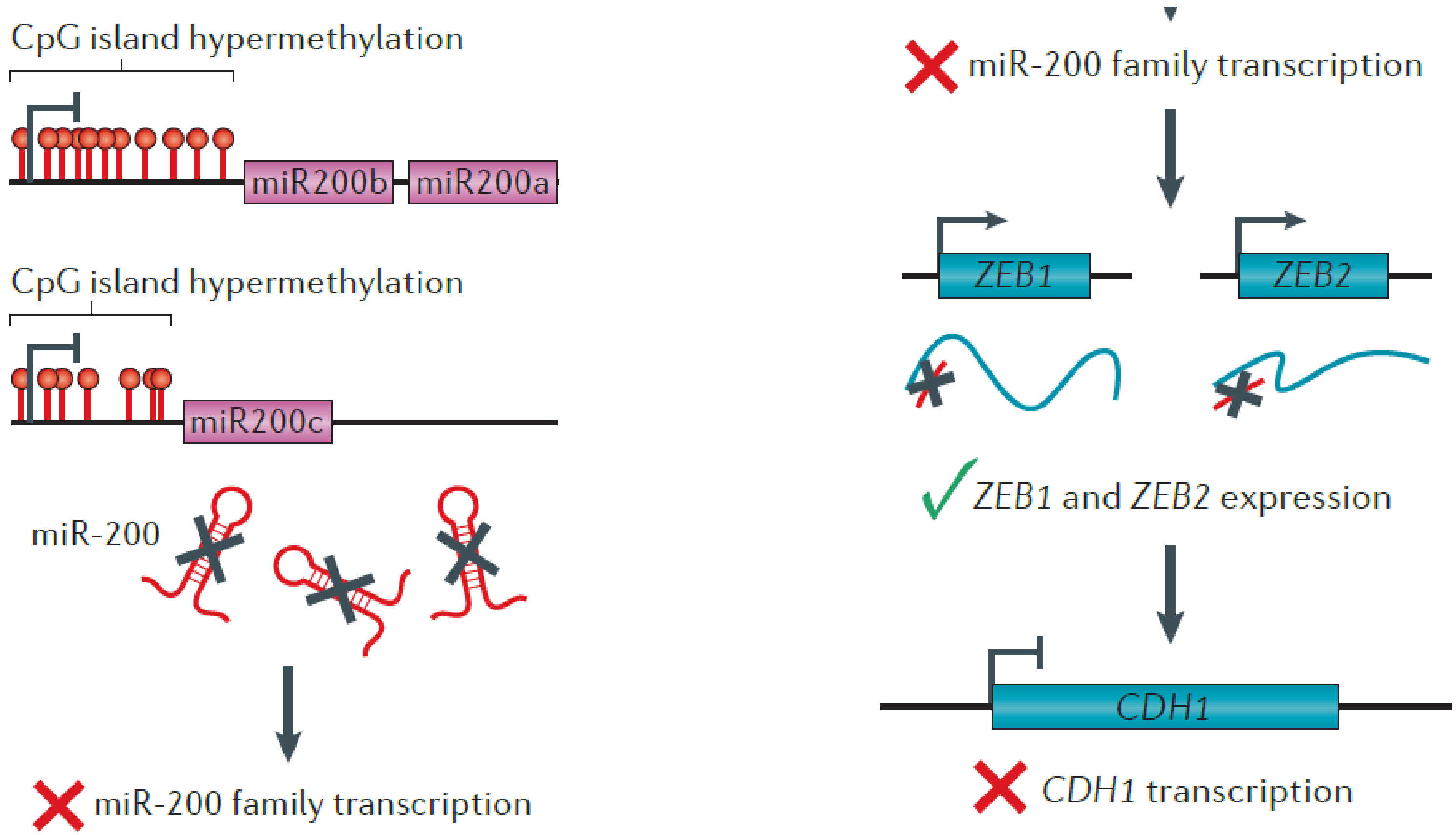
Non-coding RNAs in Alzheimer's disease



An antisense lncRNA, *BACE1-AS*, regulates the expression of the sense *BACE1* gene (labelled *BACE1-S* in the figure) through the stabilization of its mRNA. *BACE1-AS* is elevated in Alzheimer's disease, increasing the amount of BACE1 protein and, subsequently, the production of β-amyloid peptide.

Esteller (2011) *Nat Rev Genet*

Non-coding RNAs in cancer



Alterations in the epigenetic regulation of the miR-200 family are involved in epithelial-to-mesenchymal transition in cancer. Specifically, CpG island hypermethylation-associated silencing of these miRNAs in human tumours causes an upregulation of the zinc finger E-box-binding homeobox (HOX) 1 (*ZEB1*) and *ZEB2* transcriptional repressors, which, in turn, leads to a downregulation of E-cadherin *CDH1*

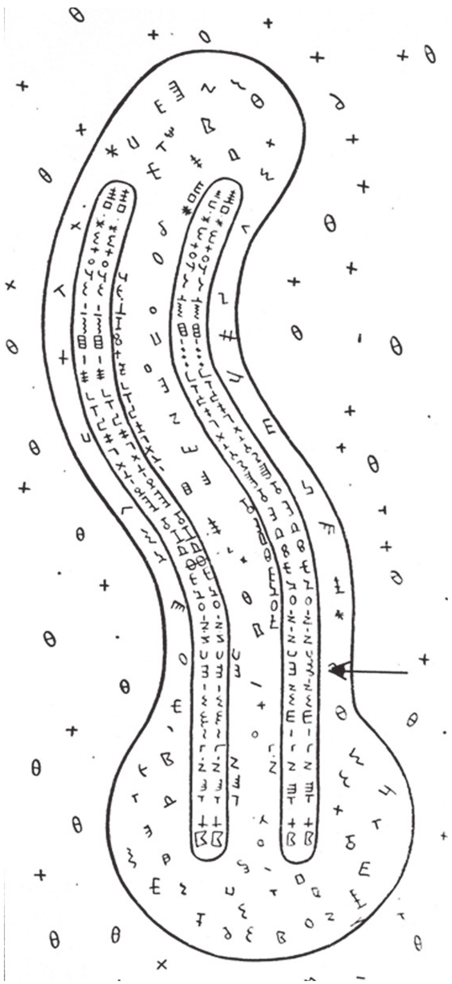
Esteller (2011) *Nat Rev Genet*

Epigenetic effects of smoking

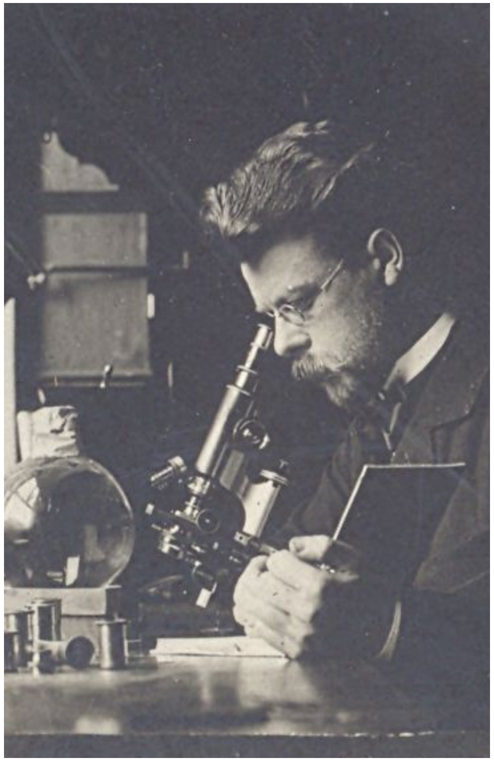
From Wikipedia, the free encyclopedia

Contents [hide]

Николай Конст. Кольцов (1872-1940)



- 1915: «Следует признать гены способными... к мутациям. Ведь во всяком органическом соединении атом водорода может быть скачкообразно заменен группой CH_3»
- 1927: *Omnis molecula ex molecula:* гипотеза о матричном воспроизведении молекул наследственности



Кольцов 1927 ⟶ Тимофеев-Ресовский, Циммер, Дельбрюк, Шредингер 1935-1945 ⟶ Уотсон, Крик 1953

Examples of coding changes in *RBFOX1*

ttttct**ag**GTTTCAAGACAACAG**ATG**AATTGTGAAAGAGAG**CAG**CTAAGG**gt**agg

M N C E R E Q L R

Synonymous change

ttttct**ag**GTTTCAAGACAACAG**ATG**AATTGTGAAAGAGAG**CAA**CTAAGG**gt**agg

M N C E R E Q L R

Non-synonymous (missense)

ttttct**ag**GTTTCAAGACAACAG**ATG**AAT**TGT**GAAAGAGAG**CAC**CTAAGG**gt**agg

M N C E R E H L R

Stop gain (nonsense)

ttttct**ag**GTTTCAAGACAACAG**ATG**AAT**TGA**GAAAGAGAGCAGCTAAGG**gt**agg

M N * E R E Q L R

Examples of coding changes in *RBFOX1*

tttct**ag**GTTTCAAGACAACAG**ATG**AATTGTGAAAGAGAG**CAG**CTAAGG**gt**agg

M N C E R E Q L R

Inframe deletion

tttct**ag**GTTTCAAGACAACAG**ATG**AATTGTGAAAGAGAG---CTAAGG**gt**agg

M N C E R E - L R

tttct**ag**GTTTCAAGACAACAG**ATG**A**AT**TGTGAAAGAGAG**CAG**CTAAGG**gt**agg

M N C E R E Q L R

Frameshift deletion

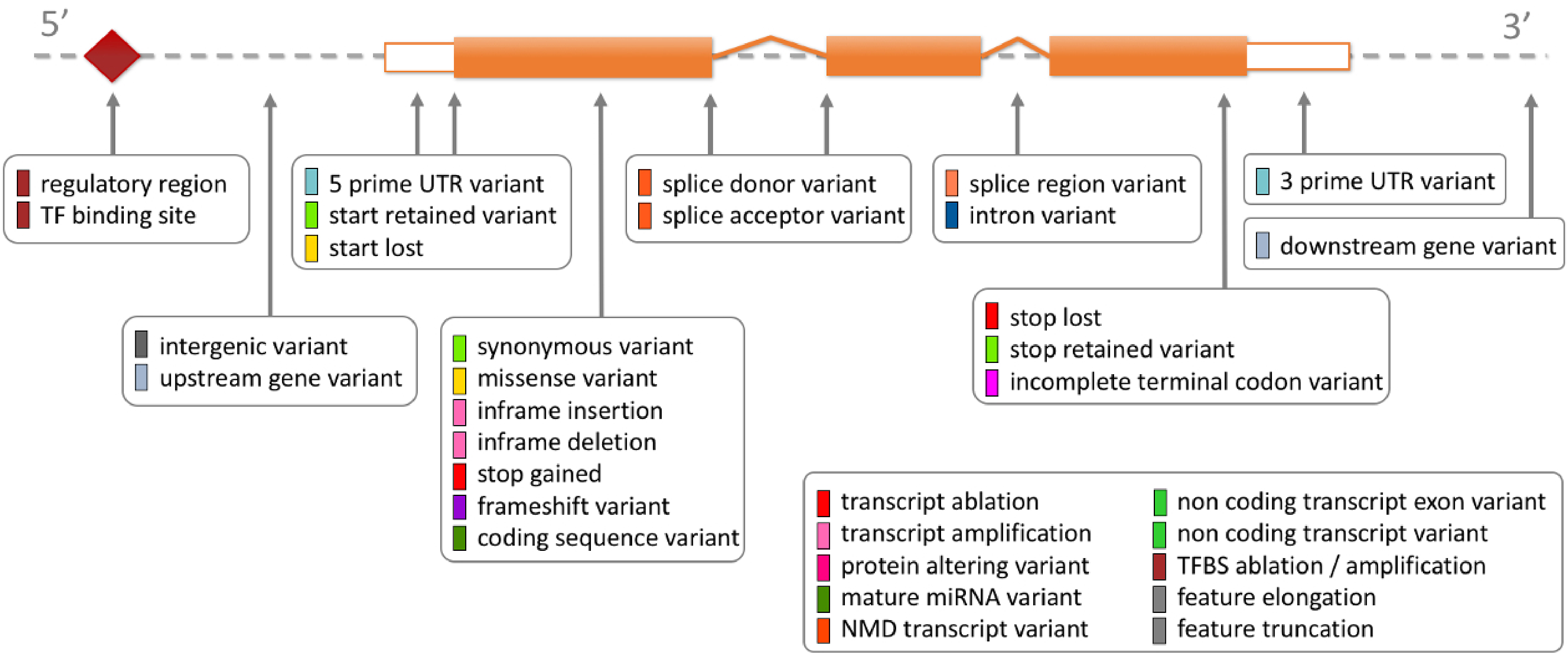
tttct**ag**GTTTCAAGACAACAG**ATG**A--TGTGAAAGAGAGCAGCTAAGG**gt**agg

M M * K R A A K

ENSEMBL Variant Effect Predictor

Variation consequences

Promoter ♦ 5’-UTR ♦ Start (ATG) ♦ Donor(GT) ♦ Acceptor(AG) ♦ … ♦ Stop(TAA,…) ♦ 3’-UTR



https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html#consequences

ENSEMBL Variant Effect Predictor

Variation consequences and impact

*	SO term	SO description	SO accession	Display term	IMPACT
	transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	Transcript ablation	HIGH
	splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO:0001574	Splice acceptor variant	HIGH
	splice_donor_variant	A splice variant that changes the 2 base region at the 5' end of an intron	SO:0001575	Splice donor variant	HIGH
	stop_gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript	SO:0001587	Stop gained	HIGH
	frameshift_variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three	SO:0001589	Frameshift variant	HIGH
	stop_lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	SO:0001578	Stop lost	HIGH
	start_lost	A codon variant that changes at least one base of the canonical start codon	SO:0002012	Start lost	HIGH
	transcript_amplification	A feature amplification of a region containing a transcript	SO:0001889	Transcript amplification	HIGH
	inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequence	SO:0001821	Inframe insertion	MODERATE
	inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequence	SO:0001822	Inframe deletion	MODERATE
	missense_variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved	SO:0001583	Missense variant	MODERATE
	protein_altering_variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence	SO:0001818	Protein altering variant	MODERATE
	splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron	SO:0001630	Splice region variant	LOW
	incomplete_terminal_codon_variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed	SO:0001626	Incomplete terminal codon variant	LOW
	start_retained_variant	A sequence variant where at least one base in the start codon is changed, but the start remains	SO:0002019	Start retained variant	LOW
	stop_retained_variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains	SO:0001567	Stop retained variant	LOW
	synonymous_variant	A sequence variant where there is no resulting change to the encoded	SO:0001819	Synonymous variant	LOW

https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html#consequences

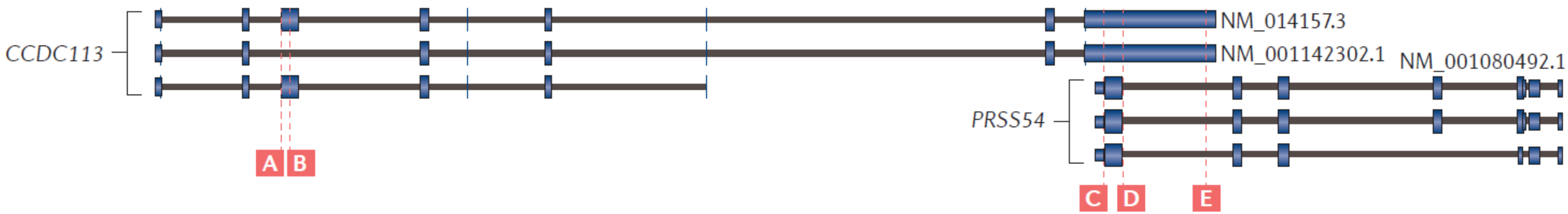
ENSEMBL Variant Effect Predictor

Variation consequences and impact

IMPACT	Consequence examples	Description
HIGH	splice_acceptor_variant, splice_donor_variant, stop_gained, stop_lost, start_lost	The variant is assumed to have high (disruptive) impact in the protein, probably causing protein truncation, loss of function or triggering nonsense mediated decay
MODERATE	inframe_insertion, inframe_deletion, missense_variant	A non-disruptive variant that might change protein effectiveness
LOW	splice_region_variant, synonymous_variant	A variant that is assumed to be mostly harmless or unlikely to change protein behaviour
MODIFIER	5_prime_UTR_variant, 3_prime_UTR_variant, intron_variant, TFBS_ablation	Usually non-coding variants or variants affecting non-coding genes, where predictions are difficult or there is no evidence of impact

https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html#consequences

Complexity of variant annotation

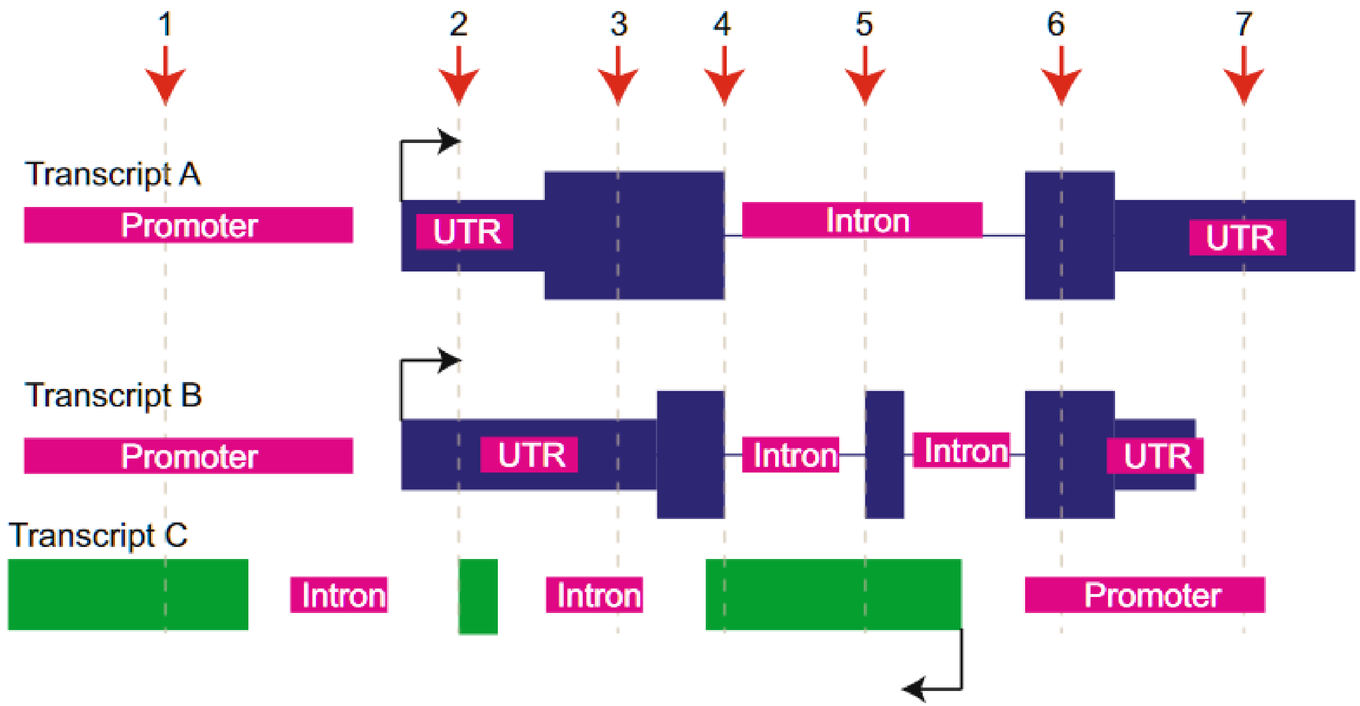


	Variant allele	Gene	Transcript change	RefSeq	Protein change	Molecular consequence
A rs765957496	G	*CCDC113*	c.228+1143A>G	NM_001142302.1	—	Intron variant
	G	*CCDC113*	c.229•2A>G	NM_014157.3	—	Splice acceptor variant
B rs775877153	A	*CCDC113*	c.228+1182T>A	NM_001142302.1	—	Intron variant
	A	*CCDC113*	c.266T>A	NM_014157.3	Met89Lys	Missense variant
C rs780162055	T	*PRSS54*	c.1135G>A	NM_001080492.1	Glu379Lys	Missense variant
	T	*CCDC113*	c.*500C>T	NM_001142302.1	—	3' UTR variant
D rs776101237	A	*PRSS54*	c.655-2A>T	NM_001080492.1	—	Splice acceptor variant
	A	*CCDC113*	c.*962T>A	NM_001142302.1	—	3' UTR variant
E rs745863465	C	*PRSS54*	c.655-18T>G	NM_001080492.1	—	Intron variant
	C	*CCDC113*	c.*996A>C	NM_001142302.1	—	3' UTR variant

A demonstration of the multiple possible effects of a single variant across transcripts and genes. The complexity of genomic annotation adds to the complexity of variant annotation. In this example, two genes, coiled-coil domain-containing 113 (*CCDC113*) and protease serine 54 (*PRSS54*) overlap on different strands of the genome, and both have multiple observed transcripts. Variants intersecting this extent of the genome show different effects depending on the gene and the transcript inspected.

Eilbeck (2017) *Nat Rev Genet*

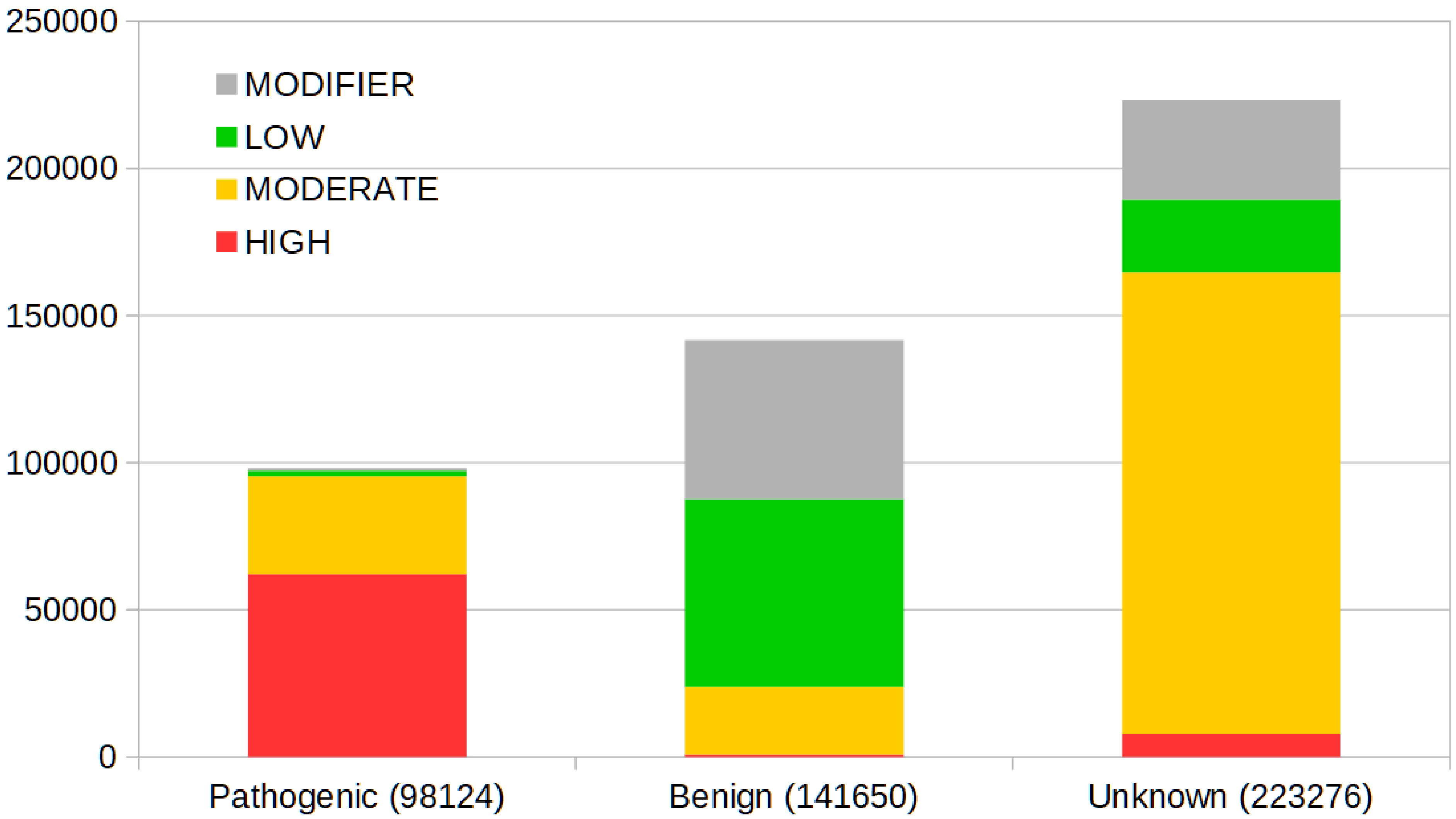
Complexity of variant annotation



Variant	Transcript A	Transcript B	Transcript C
1	Promoter	Promoter	Exon
2	Non Coding Exon	Non Coding Exon	Non Coding Splice
3	Coding Exon	Non Coding Exon	Intron
4	Coding Splice	Coding Splice	Non Coding Exon
5	Intron	Coding Splice	Non Coding Exon
6	Coding Exon	Coding Exon	Promoter
7	Non Coding Exon	Downstream	Prompter

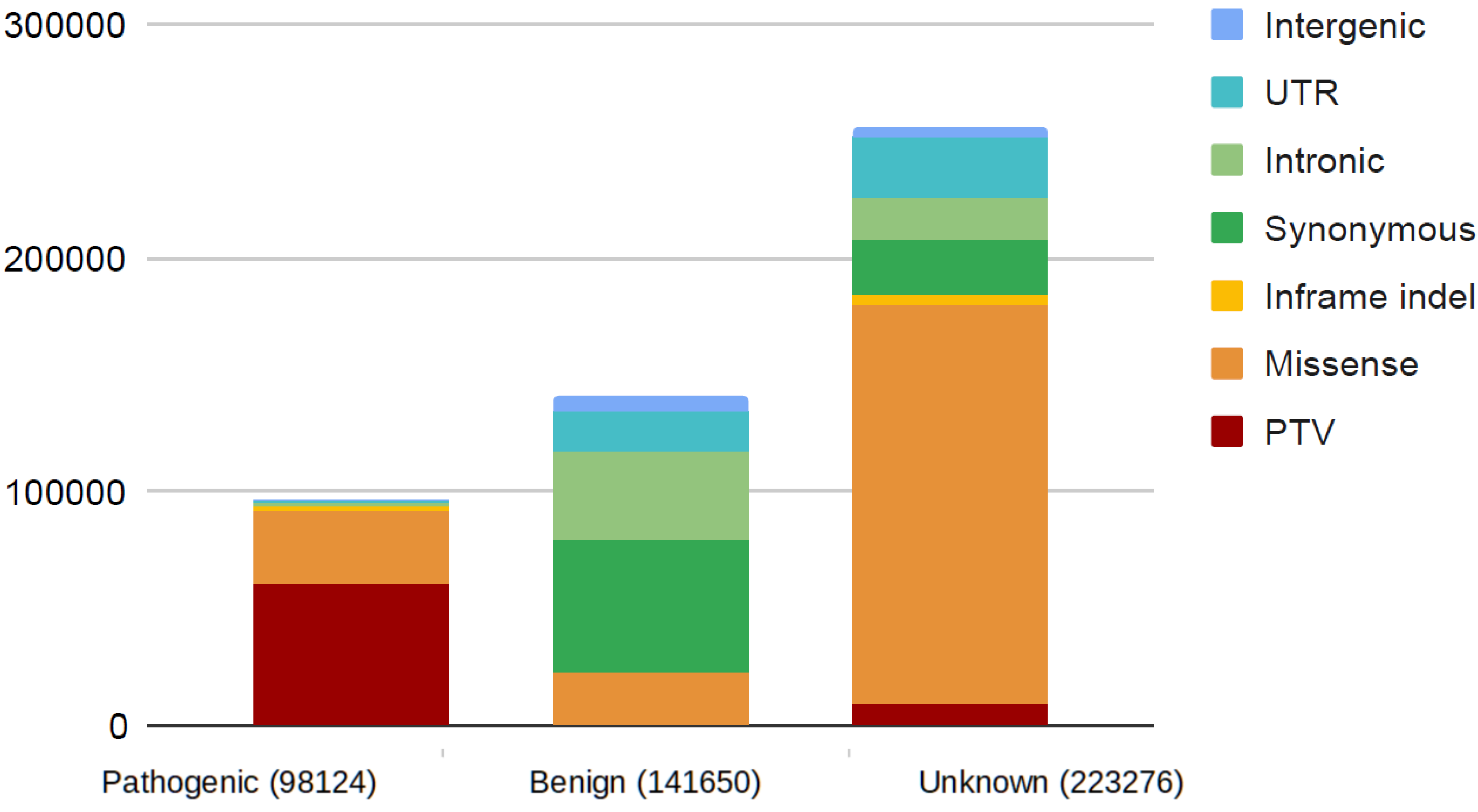
Gloss (2018) *Exp Mol Med*

Ensembl VEP annotation for ClinVar variants



ClinVar (Oct. 2019), 498,742 variants annotated with Ensembl VEP

Ensembl VEP annotation for ClinVar variants



ClinVar (Oct. 2019), 498,742 variants annotated with Ensembl VEP

Pathogenic variants in ClinVar (Oct. 2019)

Gene	Frameshift	Stop gain or loss	Splice site	Missense	Inframe	Synonymous	UTR	Intronic	Upstream	Start codon	Phenotype
HBB	30	14	21	35	3	1	7	12	7	4	Beta thalassemia
LDLR	387	171	51	77	9	3	7	6	0	2	Familial hypercholesterolemia
CFTR	123	111	70	105	5	3	0	20	0	4	Cystic fibrosis
GALT	21	15	11	100	1	2	0	4	1	1	Deficiency of UDPglucose-hexose-1-phosphate uridylyltransferase
KCNQ2	61	20	20	102	7	2	0	1	1	1	Benign familial neonatal seizures; Early infantile epileptic encephalopathy
MECP2	268	60	12	27	12	2	0	1	0	3	Mental retardation; Rett syndrome
MLH1	316	132	76	69	4	6	1	11	0	10	Hereditary nonpolyposis colon cancer; Lynch syndrome
OTC	22	32	39	203	5	2	0	7	0	4	Ornithine carbamoyltransferase deficiency

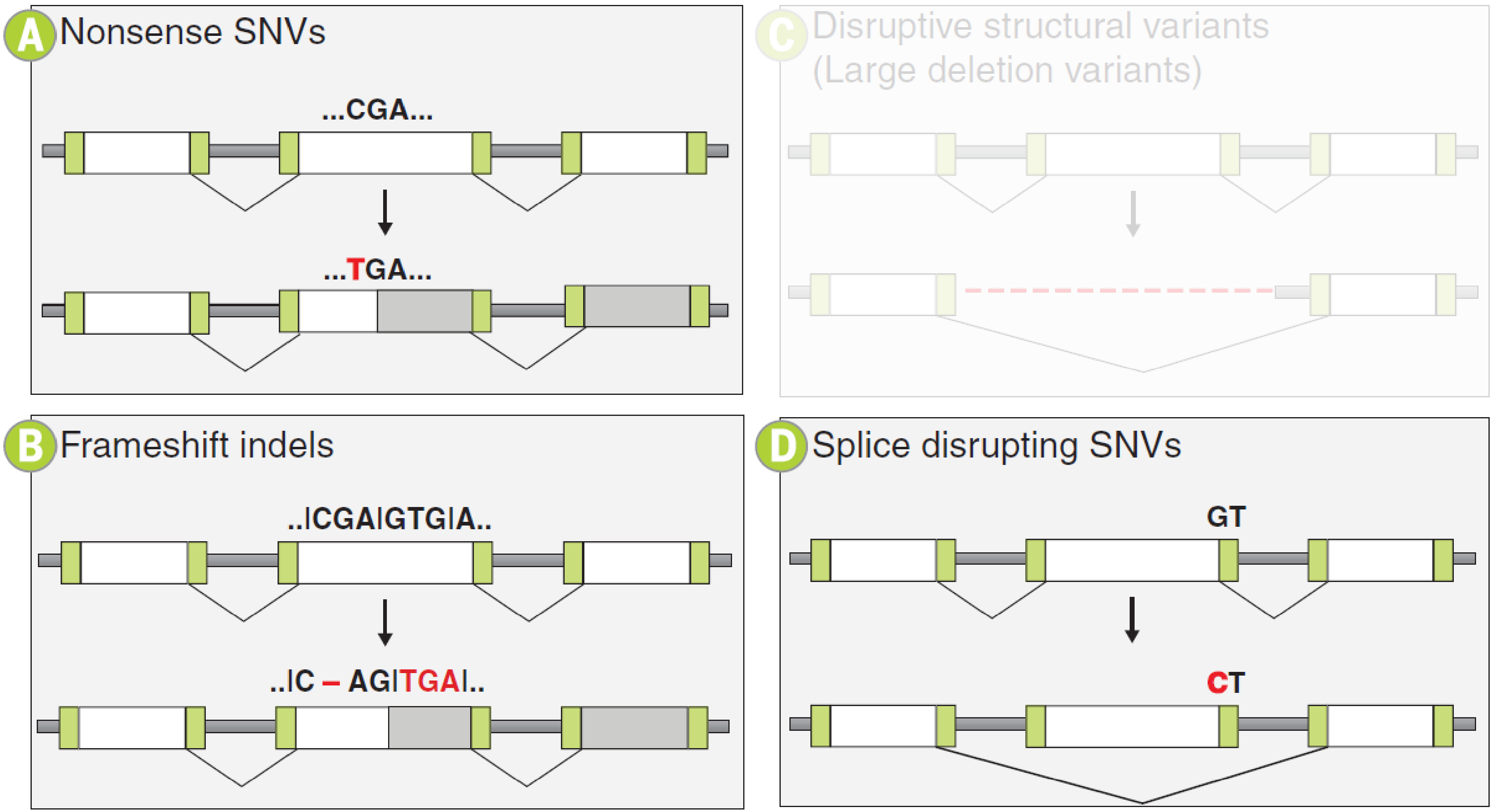
Exercise

Use ClinVar (OMIM) to find and save one example of disease-associated pathogenic mutation for *each* annotation type:

- stop-gain
- synonymous
- missense
- splice-site
- frameshift indel

PTVs and LoF variants

Protein-truncating variants: stop-gain, splice site, frameshift indels.
VEP impact: HIGH.



Rivas (2015) *Science*

PTVs and LoF variants

Protein-truncating variants: stop-gain, splice site, frameshift indels. VEP impact: HIGH. *However, not all PTVs are loss-of-function*

LOFTEE tool (K.Karczewski et al): filters and flags to predict pLoF (putative LoF) from candidate PTVs. https://github.com/konradjk/loftee

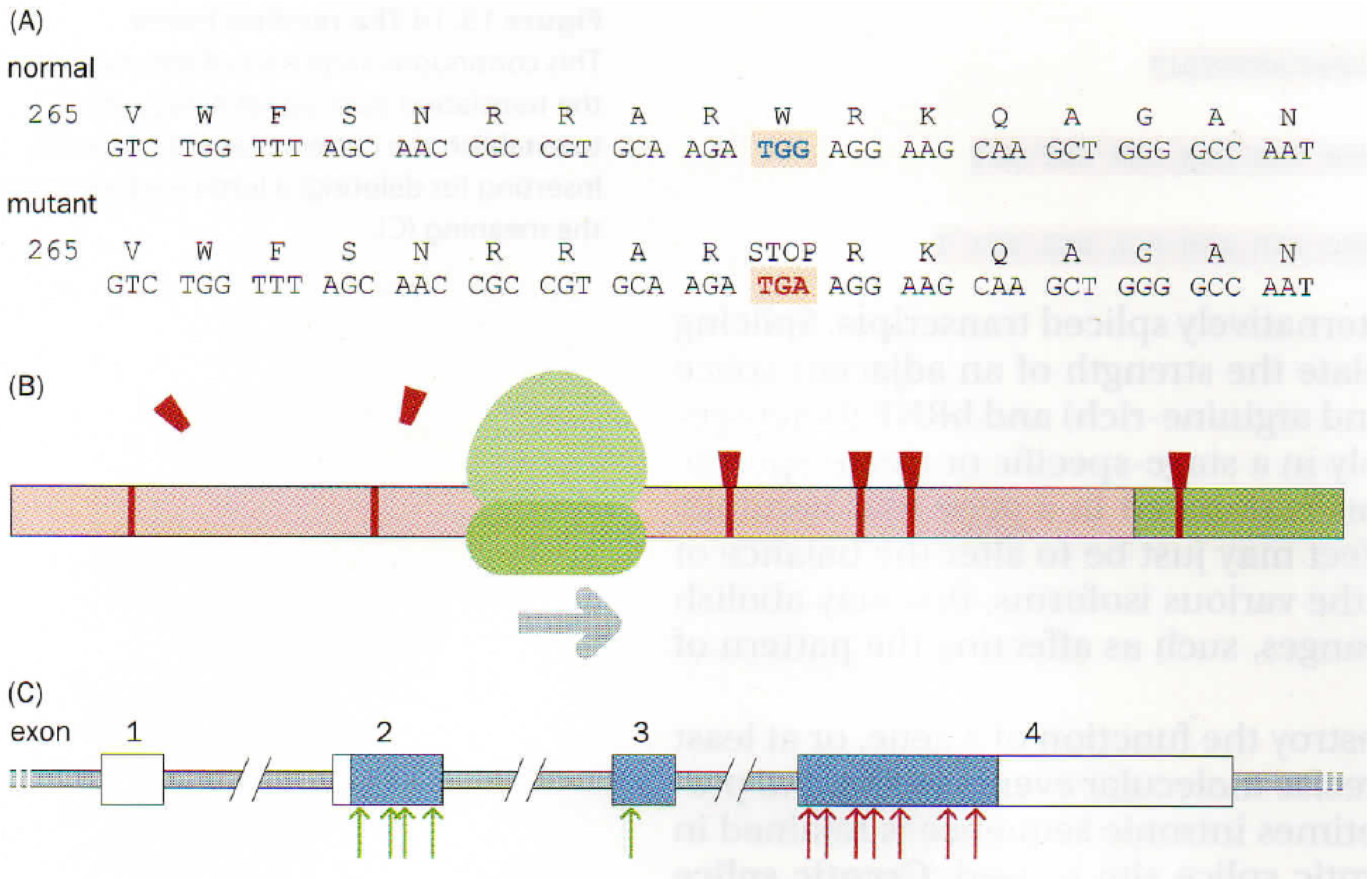
PTVs not predicted as pLoF, examples:

- Stop-gain and frameshift variants near the end of the transcript, based on the 50 bp rule
- Variants in an exon with non-canonical splice sites (GT, AG) around it
- Splice site variants rescued by nearby, in-frame splice site
- Variants in small introns

Flagged PTVs, examples:

- Variants in NAGNAG sites (acceptor sites rescued by in-frame acceptor site)
- Variants that fall in an intron with a non-canonical splice site

PTVs and nonsense-mediated decay (NMD)



(A) G>A change in exon 6 of the *PAX3* gene (B) Nonsense-mediated decay (NMD). Splice junctions (red bars) retain proteins of the exon junction complex (EJC, red triangles). Ribosome moves along the mRN A and displaces the EJC proteins. If it encounters a premature stop codon and detaches before displacing all EJCs, the mRNA is targeted for degradation. **Stop codons in the last exon or less than 50 nucleotides upstream of the last splice junction (the green zone) do not trigger NMD.** (C) Depending on whether or not a premature stop codon triggers NMD, the consequences of a nonsense mutation can be very different.

Strachan, Read – *Human Molecular Genetics*

PTVs and nonsense-mediated decay (NMD)

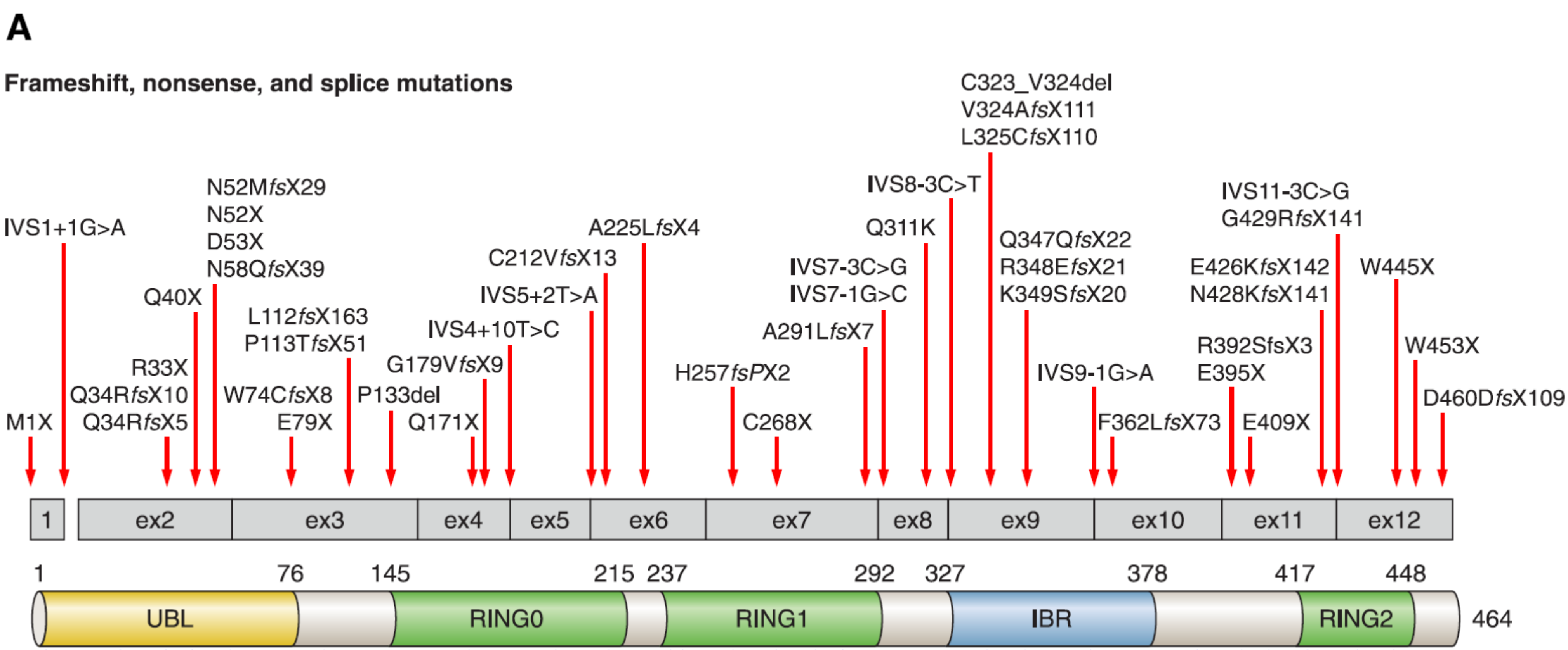
Ideally: **PTV → NMD → Transcript level → Protein level → Cellular functions**

However, variation in mRNA and protein expression levels are often uncorrelated: the reduction in RNA levels may not reduce the protein level, and vice versa

Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K., and Gilad, Y. (2015). Impact of Regulatory Variation from RNA to Protein. Science 347, 664–667.

Narasimhan VM, Xue Y, Tyler-Smith C. Human Knockout Carriers: Dead, Diseased, Healthy, or Improved? Trends in Molecular Medicine. 2016;22(4):341-351. doi:10.1016/j.molmed.2016.02.006.

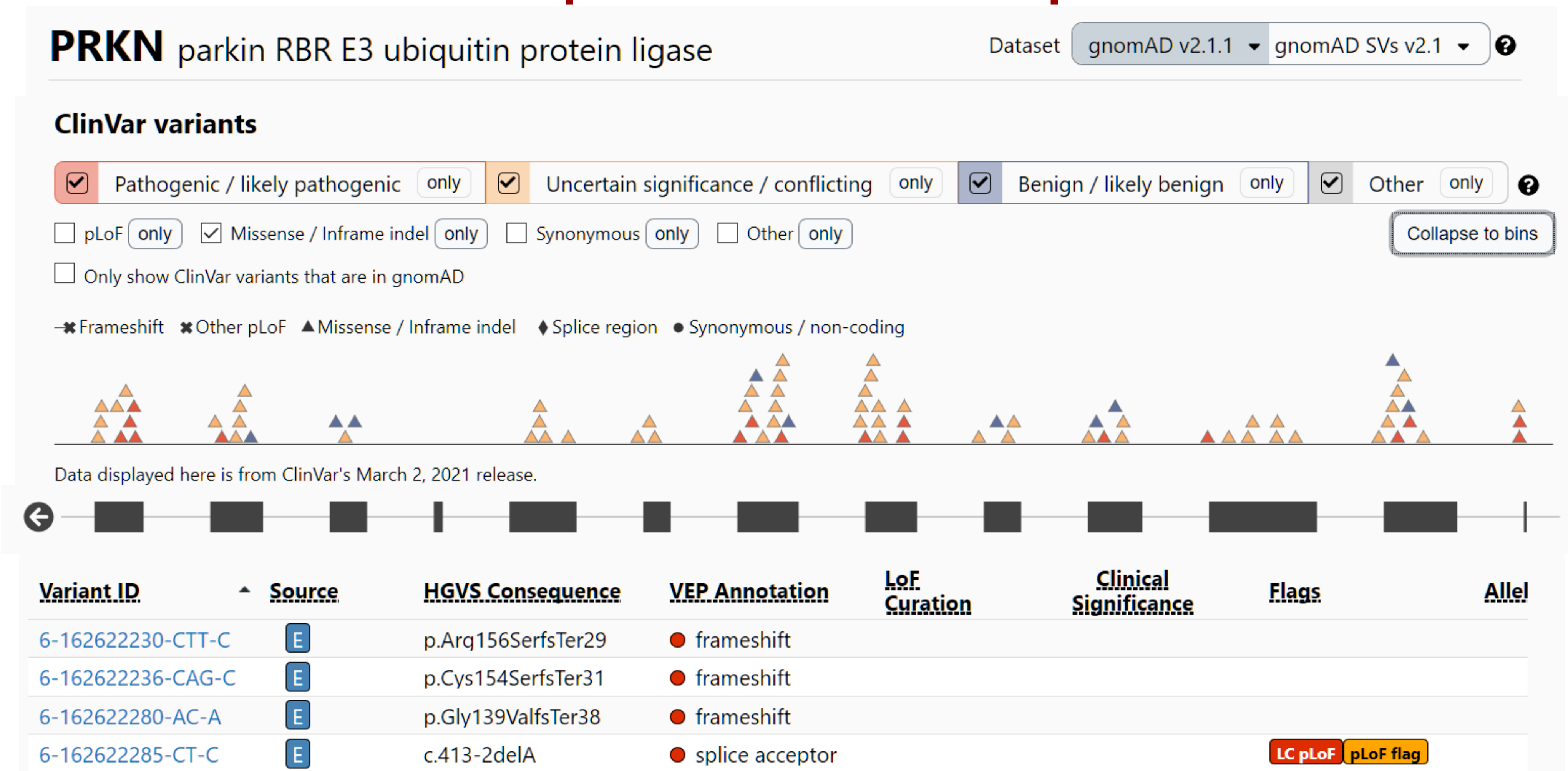
Examples of PTV impact



Mutations in the Parkin RBR E3 Ubiquitin Protein Ligase *PRKN* are the most frequent known cause of early-onset (40–50 yr) Parkinson's disease. PD is the second most common neurodegenerative disorder, after Alzheimer's disease, with prevalence in industrialized countries ~0.3%.

Corti (2011) *Physiol Rev*

Examples of PTV impact



Mutations in the Parkin RBR E3 Ubiquitin Protein Ligase *PRKN* are the most frequent known cause of early-onset (40–50 yr) Parkinson’s disease. PD is the second most common neurodegenerative disorder, after Alzheimer’s disease, with prevalence in industrialized countries ~0.3%.

Corti (2011) *Physiol Rev*

Examples of PTV impact

Protein-truncating variants: stop-gain, splice site, frameshift indels.
VEP impact: HIGH.

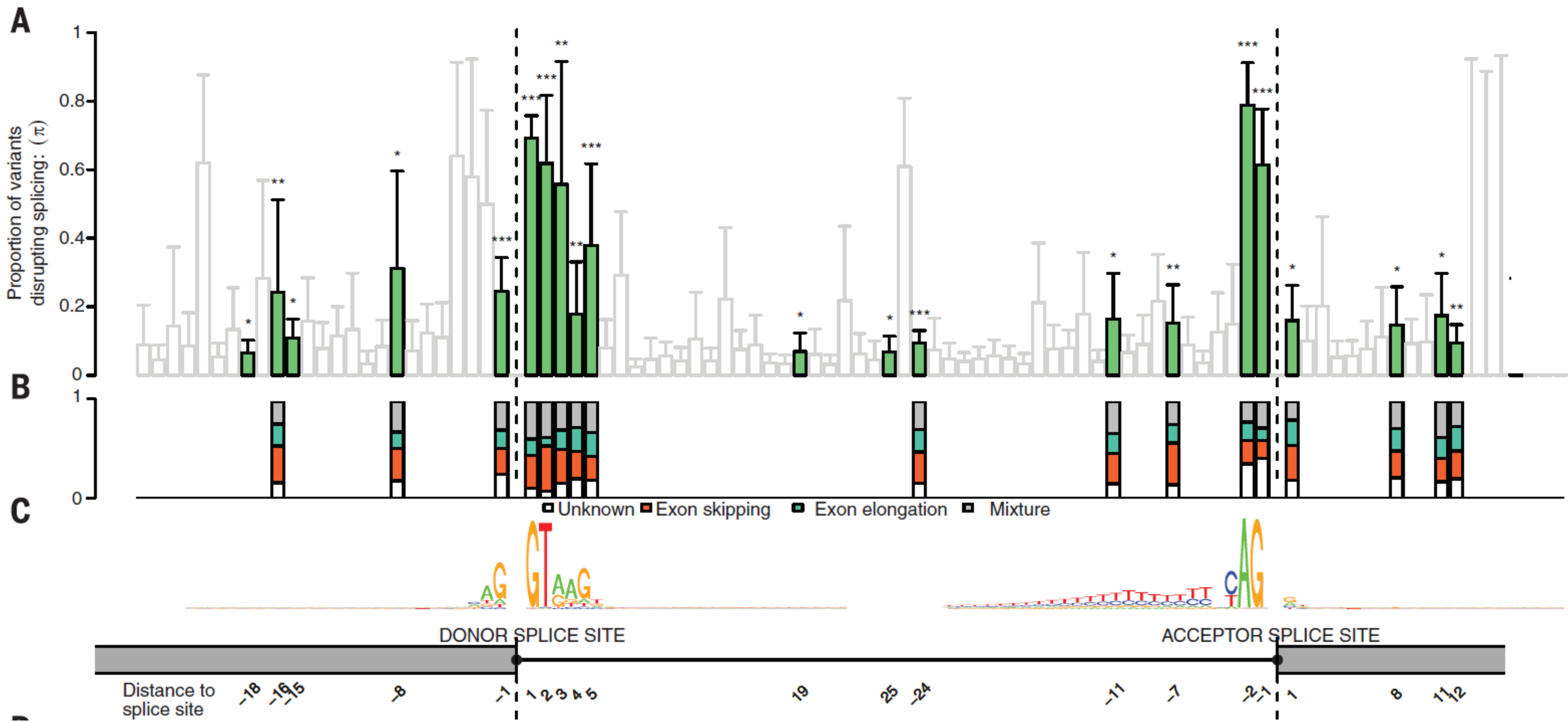


Fig. 3. Splicing disruption. (A) Proportion of variants disrupting splicing at each distance +/- 25 bp from donor and acceptor site (B) Classification of splice disruption events: exon skipping, exon elongation and mixture (C) Diagram of donor and acceptor splice junctions and sequence logo of represented sequences.

Rivas (2015) *Science*

Examples of PTV impact

1. Narasimhan VM, Xue Y, Tyler-Smith C. (2016) Human Knockout Carriers: Dead, Diseased, Healthy, or Improved? *Trends Mol Med* 22:341-351.

- A knockout of the immune gene *IRF7* was shown to confer **susceptibility to flu viruses**, leading to life-threatening influenza in an otherwise healthy child (Ciancanelli 2015 *Science*)
- Instances where a naturally-occurring **LoF variant proves beneficial to health**. These discoveries have stimulated drug development:
  - lowering LDL levels: *PCSK9*
  - decreasing susceptibility to HIV: *CCR5*
  - increasing endurance: *ACTN3*
  - increasing sepsis resistance: *CASP12*
  - reduced triglyceride levels in humans: *APOC3*

2. DeBoever, C., Tanigawa, Y., Lindholm, M.E., et al. (2018). Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. *Nat Commun* 9, 1–10.

- 18,228 PTVs × 135 phenotypes; find **27 associations between medical phenotypes and PTVs** in genes outside the MHC

Examples of PTV impact

1. The stop-gain variant in *GNAS* (MIM:139320) is present in the highly variable **first exon** of the gene and is likely to result in nonsense-mediated RNA decay; in contrast, pathogenic *GNAS* variants that cause Albright hereditary osteodystrophy (MIM:103580) are located in **later**, highly constrained exons.

2. Similarly, the stop-gain variant in *TGIF1* (MIM:602630) is located in the **first exon**, where multiple PTVs in gnomAD are also located, but *TGIF1* pathogenic variants causing holoprosencephaly are located in the **final exons**, where they affect DNA binding affinity.

3. Finally, a frameshift deletion in *HIST1H1E* (MIM:142220) is located near **the start** of the single exon of this gene; however, pathogenic *HIST1H1E* frameshift deletions that cause child overgrowth and intellectual disability are located near **the end** of the exon, where they result in a truncated histone protein with lower net charge that is less effective at binding DNA.

We believe that these three rare PTVs are benign because of their locations, despite the fact that they occur in genes that cause dominant DD via haploinsufficiency.

Wright (2019) *Am J Hum Genet*

Missense variant, classic example

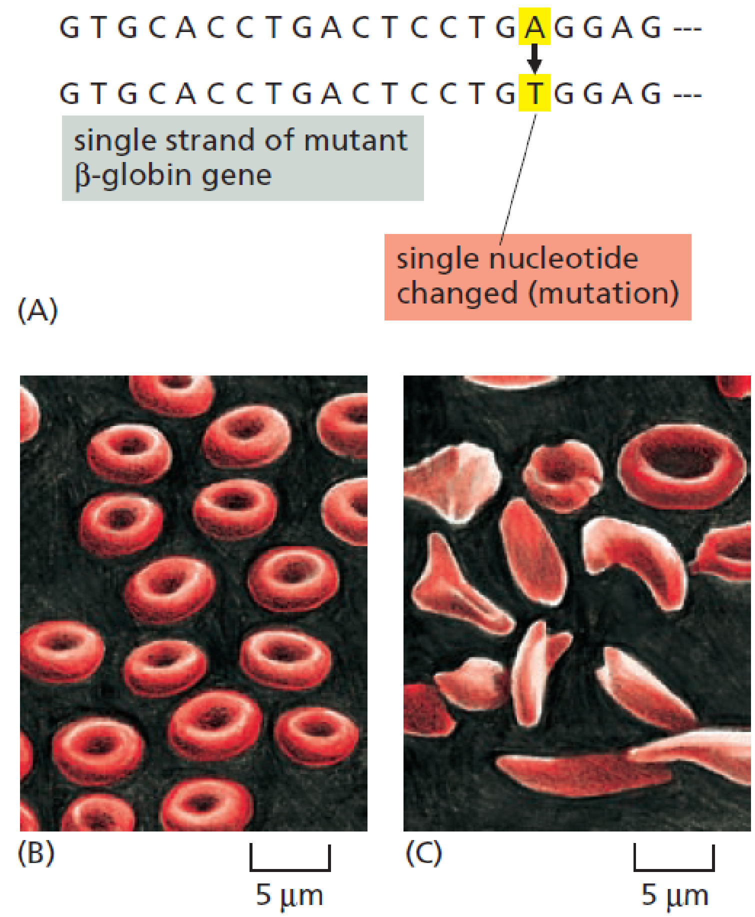
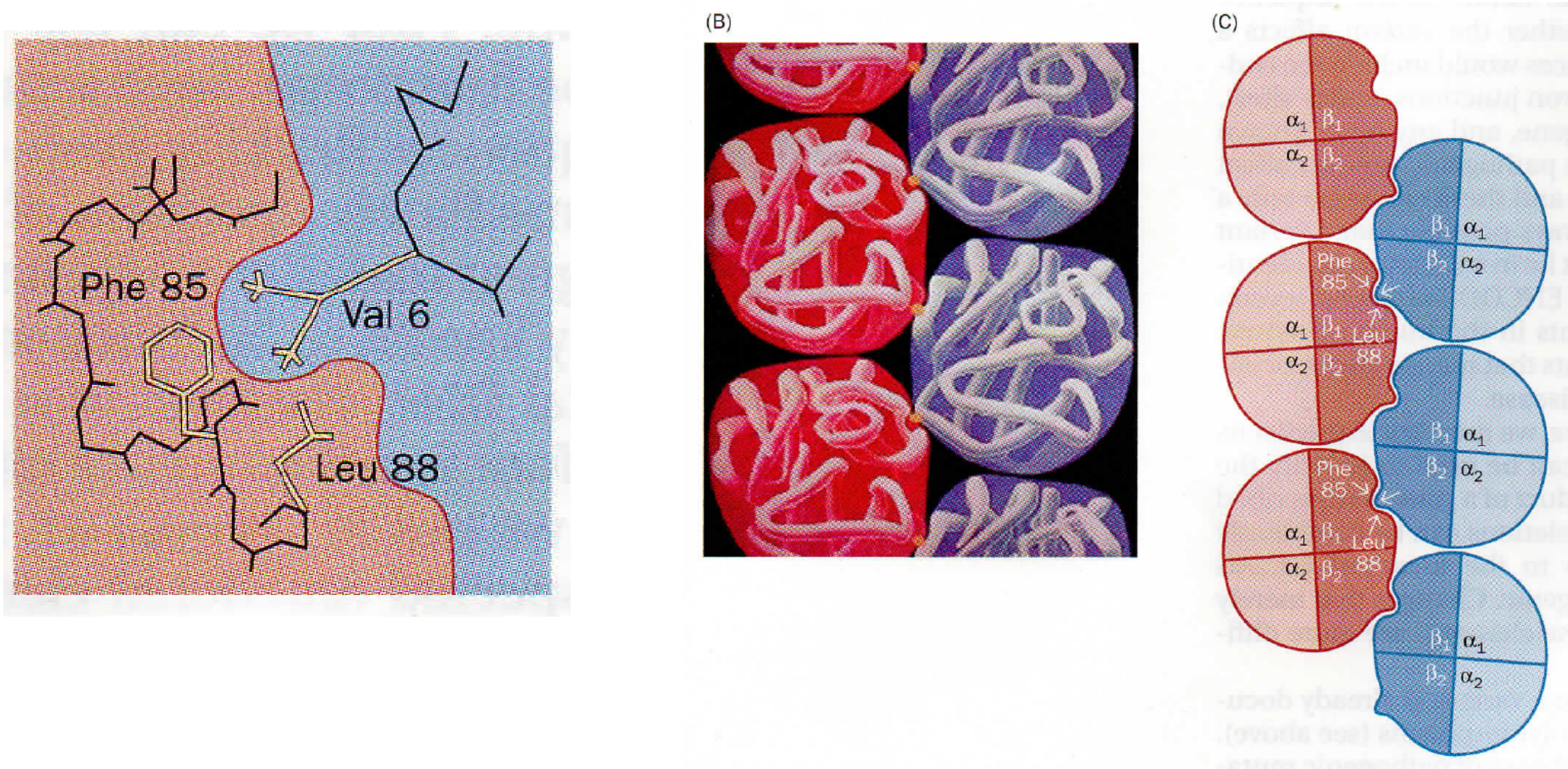


Figure 6–19 **A single nucleotide change causes the disease sickle-cell anemia.** (A) β-globin is one of the two types of subunit that form hemoglobin (see Figure 4–20). A single nucleotide change (mutation) in the β-globin gene produces a β-globin subunit that differs from normal β-globin only by a change from glutamic acid to valine at the sixth amino acid position. (Only a small portion of the gene is shown here; the β-globin subunit contains a total of 146 amino acids.) Humans carry two copies of each gene (one inherited from each parent); a sickle-cell mutation in one of the two β-globin genes generally causes no harm to the individual, as it is compensated for by the normal gene. However, an individual who inherits two copies of the mutant β-globin gene displays the symptoms of sickle-cell anemia. Normal red blood cells are shown in (B), and those from an individual suffering from sickle-cell anemia in (C). Although sickle-cell anemia can be a life-threatening disease, the mutation responsible can also be beneficial. People with the disease, or those who carry one normal gene and one sickle-cell gene, are more resistant to malaria than unaffected individuals, because the parasite that causes malaria grows poorly in red blood cells that contain the sickle-cell form of hemoglobin.

HBB.Glu7Val **Sickle cell anemia** [MIM:603903]: Characterized by abnormally shaped red cells resulting in **chronic anemia and periodic episodes of pain, serious infections and damage to vital organs**. Normal red blood cells are round and flexible and flow easily through blood vessels, but in sickle cell anemia, the abnormal hemoglobin (called Hb S) causes red blood cells to become stiff. They are C-shaped and resembles a sickle. These stiffer red blood cells can led to microvascular occlusion thus cutting off the blood supply to nearby tissues // www.genecards.org

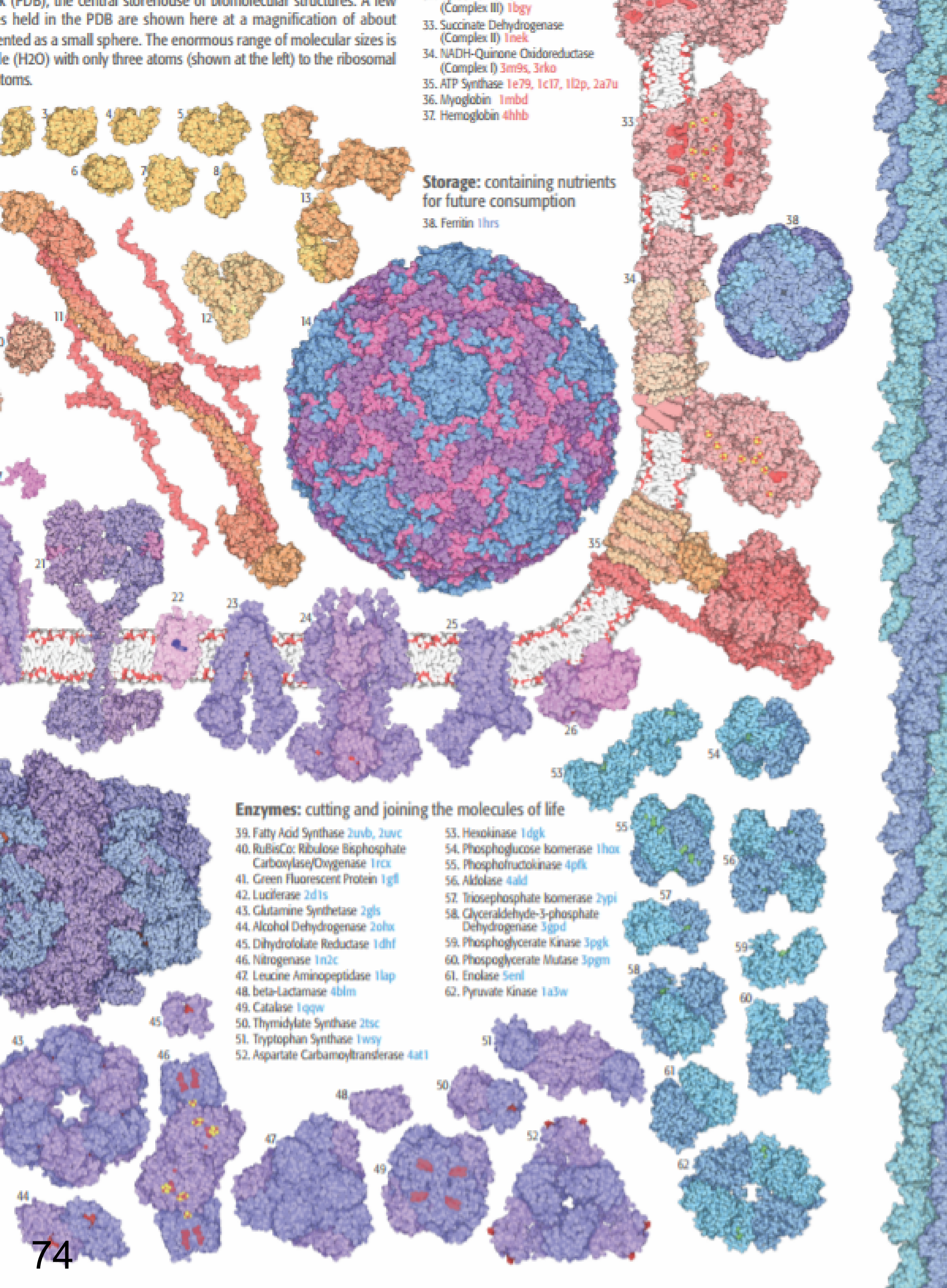
Alberts - *Essential Cell Biology*

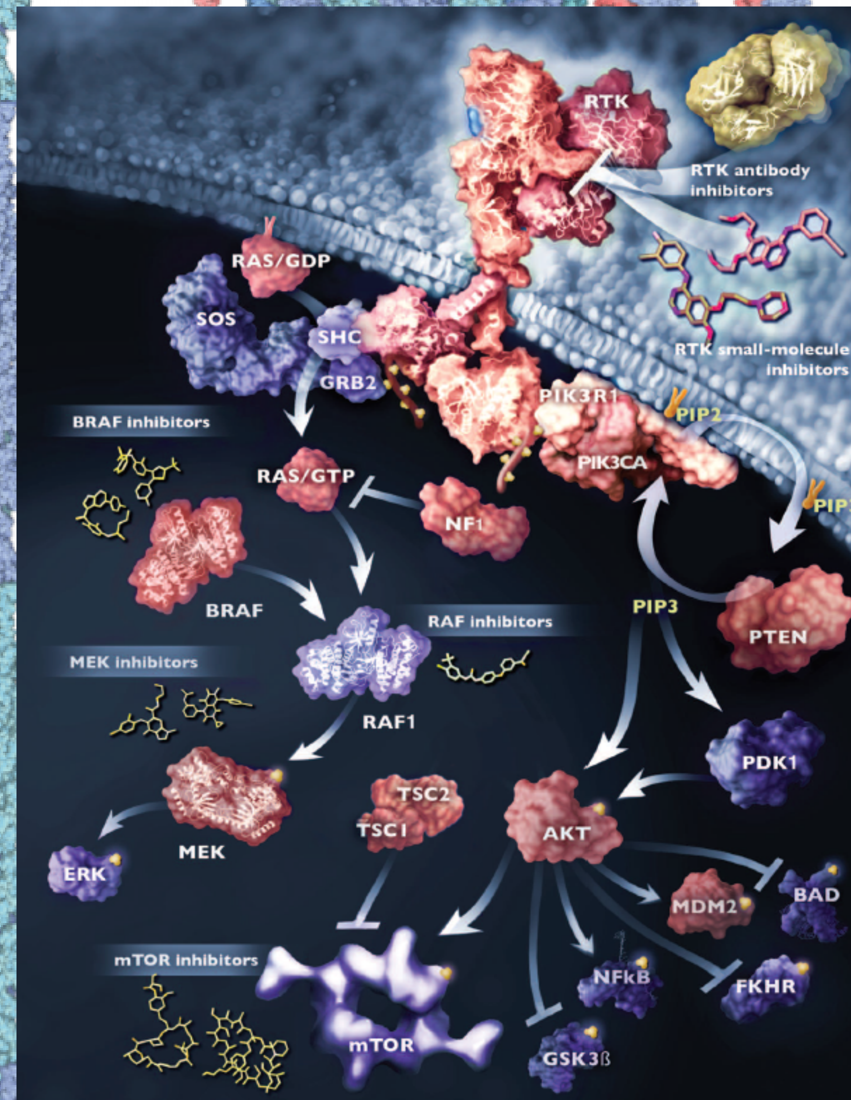
Missense variant, classic example



The sickle cell mutation. An A>T mutation in the β-globin *(HBB)* gene causes an amino acid change in the β-globin protein. The mutation replaces glutamic acid, a hydrophilic charged amino acid, with valine, a hydrophobic nonpolar amino acid. This change on the surface of the globin protein allows adhesive interactions between hemoglobin molecules.

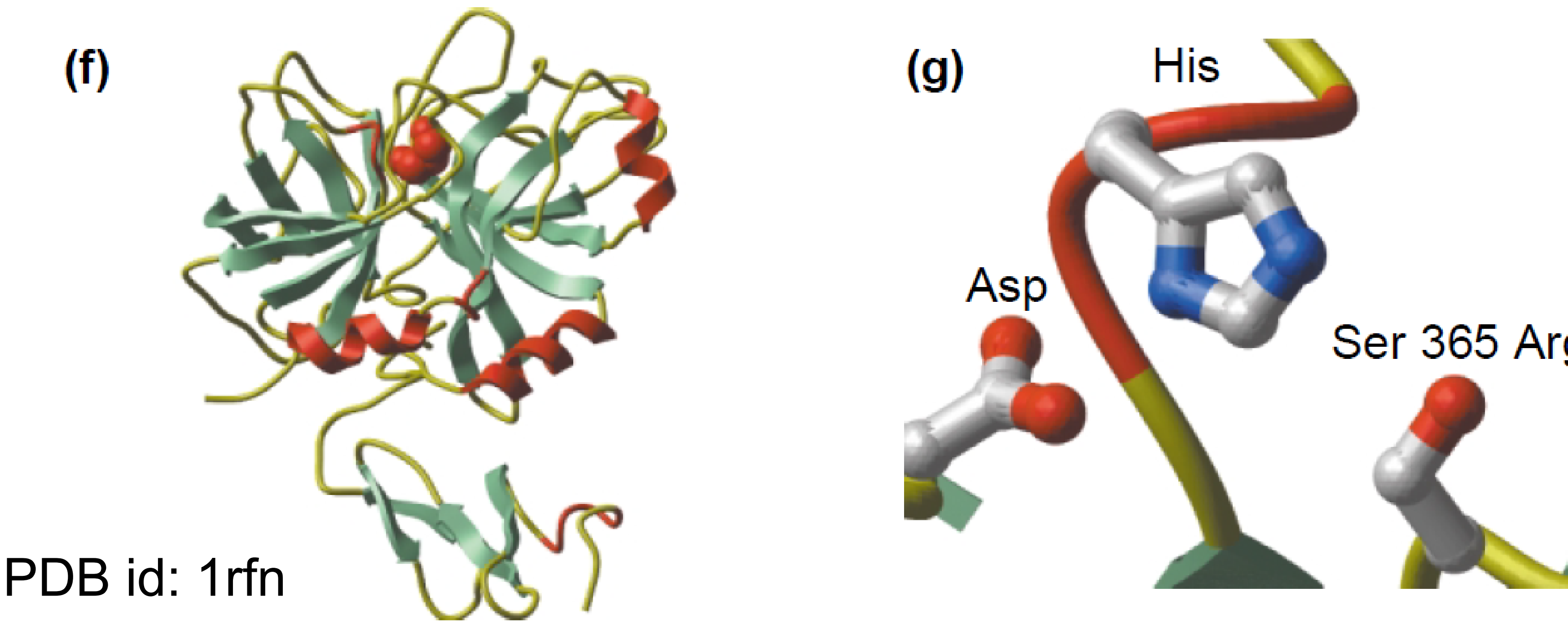
Strachan, Read – *Human Molecular Genetics*





Vogelstein (2013) *Science*
Protein Data Bank rcsb.org

Examples of missense disease variants



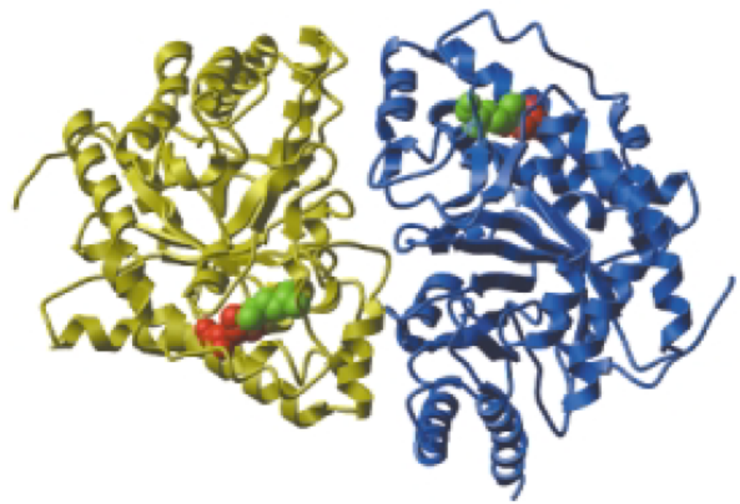
Factor IX *F9* is a serine protease with Ser-His-Asp catalytic triade that participates in the intrinsic pathway of blood coagulation by converting factor X to its active form Xa. Disease mutations in *F9* are associated with the X-linked recessive bleeding disorder haemophilia B (OMIM:306900).

Disruption of catalytic residues. Mutations of the catalytic serine residue to an arginine results in the loss of enzyme activity and a severe haemophilia phenotype.

Steward (2003) *Trends Genet*

Examples of missense disease variants

Introduction of buried charged residues:
Met165Arg ⟹ arginine sidechain cannot be accommodated in a hydrophobic pocket ⟹ no soluble protein.



PDB id: 1uro

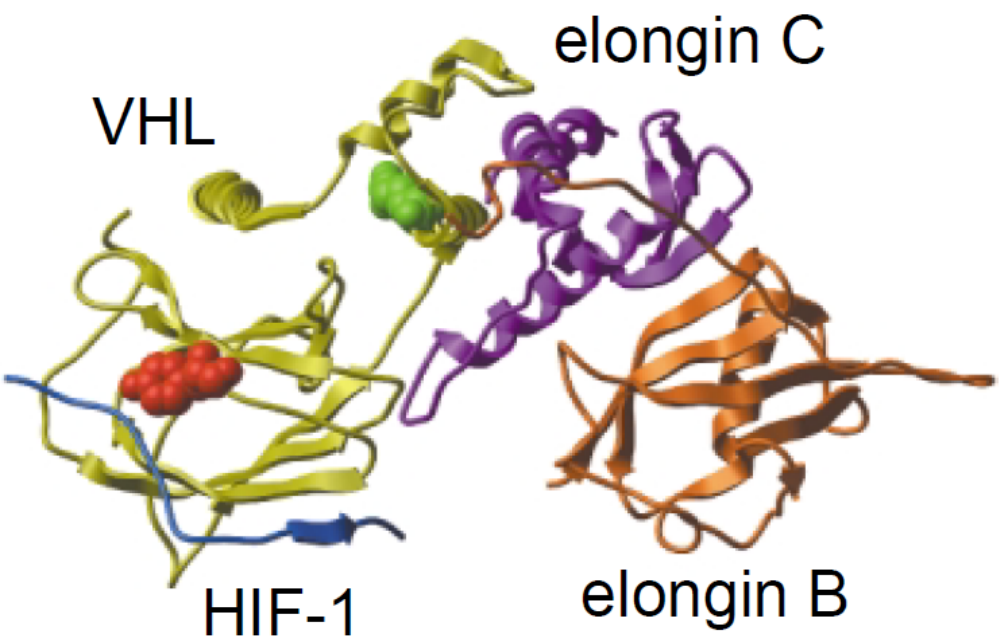
Size changes in the hydrophobic core:
Leu195Phe ⟹ rearrangement of surrounding side-chains ⟹ 30% of the wild-type activity.

Mutations in the uroporphyrinogen decarboxylase *UROD* are associated with Porphyria cutanea tarda (OMIM:176100), accumulation of uroporphyrins in the liver and plasma, leading to skin fragility and photosensitive dermatitis.

Steward (2003) *Trends Genet*

Examples of missense disease variants

Disruption of protein–protein interactions:
Tyr98His destroys binding between HIF and VHL ⟹ HIF not degraded ⟹ over-expression of angiogenic growth factors ⟹ local proliferation of blood vessels.



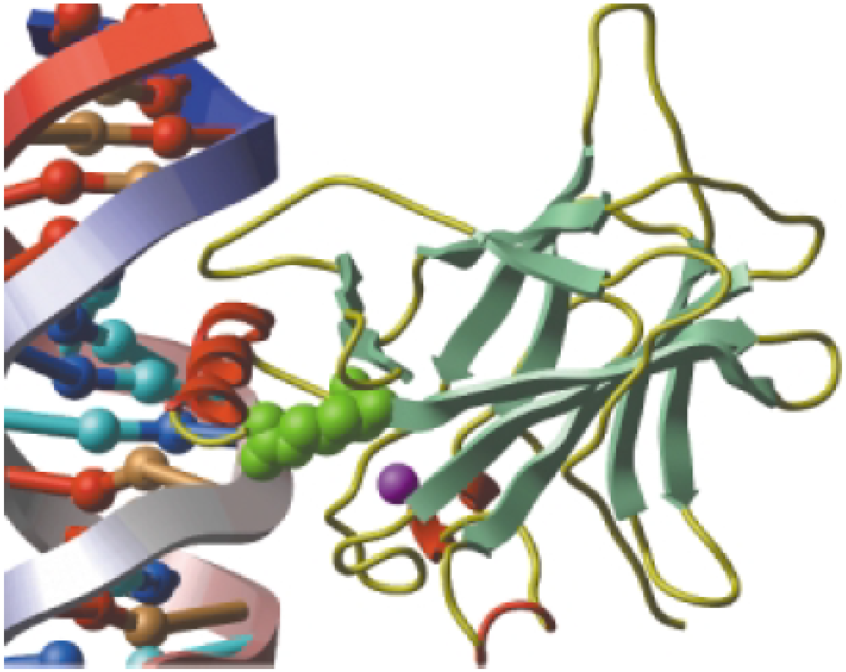
PDB id: 1lm8

Von Hippel-Lindau syndrome (OMIM:193300) is an inherited predisposition to a variety of cancers. Von Hippel-Lindau disease tumor suppressor *VHL* codes for a protein with two structural domains. The β-domain of VHL binds to hypoxia-inducible transcription factor HIF, ultimately leading to HIF degradation.

Steward (2003) *Trends Genet*

Examples of missense disease variants

Disruption of DNA binding
Arg273 contacts the DNA phosphate backbone with its charged side-chain. Arg273His is associated with low p53 DNA-binding and Li-Fraumeni syndrome.



PDB id: 1tsr

Li-Fraumeni syndrome (OMIM 191170), a predisposition to a broad spectrum of cancers at an early age. Cellular tumor antigen p53 (*TP53)* is a tumor suppressor in many tumor types, induces growth arrest or apoptosis. Three functional domains: an N-terminal transcription factor domain, a DNA-binding core domain, and a C-terminal homooligomerization domain.

Steward (2003) *Trends Genet*

Examples of missense disease variants

Fig. 5 Protein structural variation. Organization of the descriptive VariO terms, which facilitate very detailed annotation of observed effects

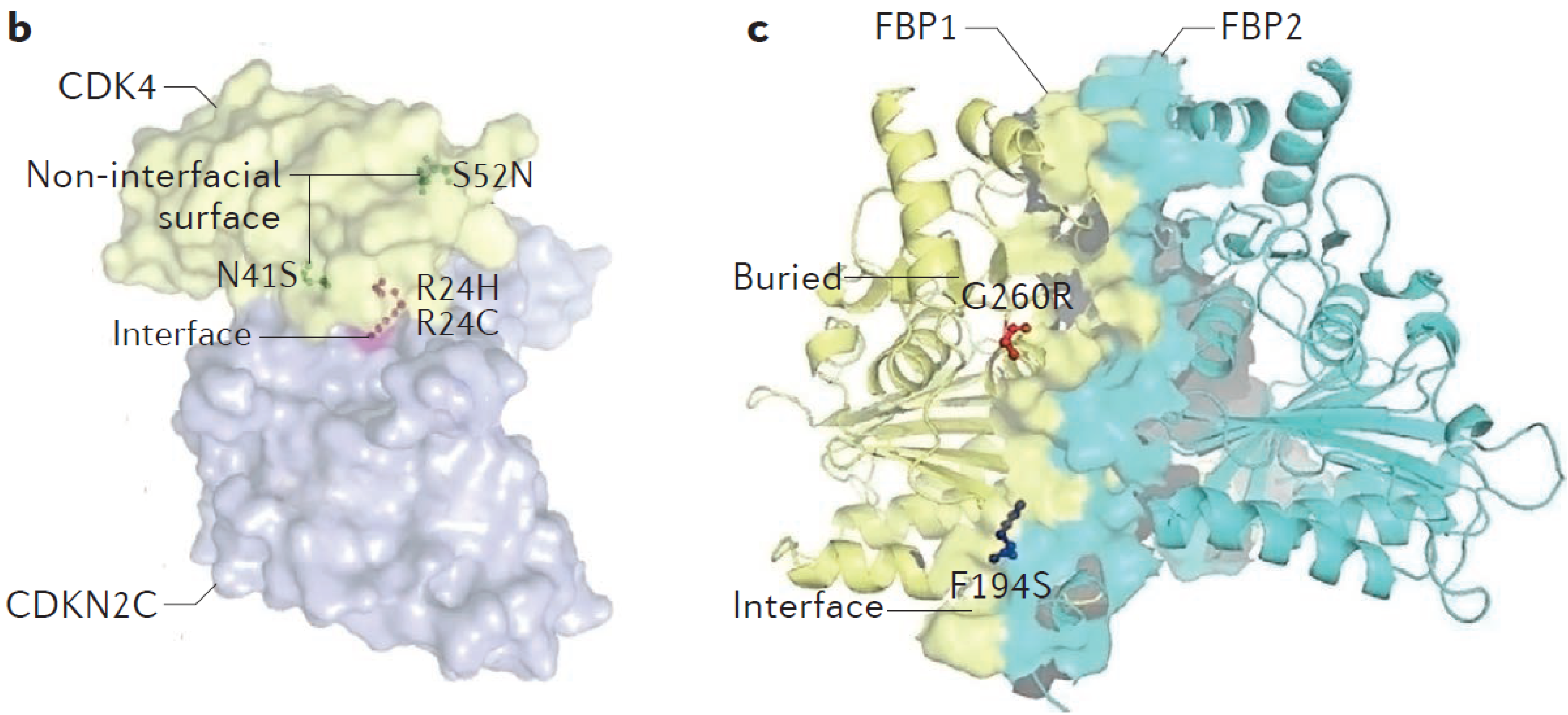
VariO: Variant effect on protein…

- Dynamics
- Quaternary structure
- Amino acid size
- Folding rate
- Interactions
- Post-translational modification
- Secondary structural element
- Fold
- Epigenetic modification
- Abundance
- Accessibility
- Activity
- Charge
- Degradation
- Solubility
- Stability
- Subcellular localization
- …

www.variationontology.org

Vihinen (2015) *Human Genet*

Missense disease mutations: stability or PPI?



b | Locations of residues affected by mutations are highlighted on the cyclin-dependent kinase 4 (CDK4) structure based on homology modelling (PDB: 1bi7). CDKN2C, CDK inhibitor 2C. **c** | Locations of residues affected by mutations are highlighted on the fructose bisphosphatase 1 (FBP1) structure (PDB: 1fpi).

Yi (2017) *Nat Rev Genet*

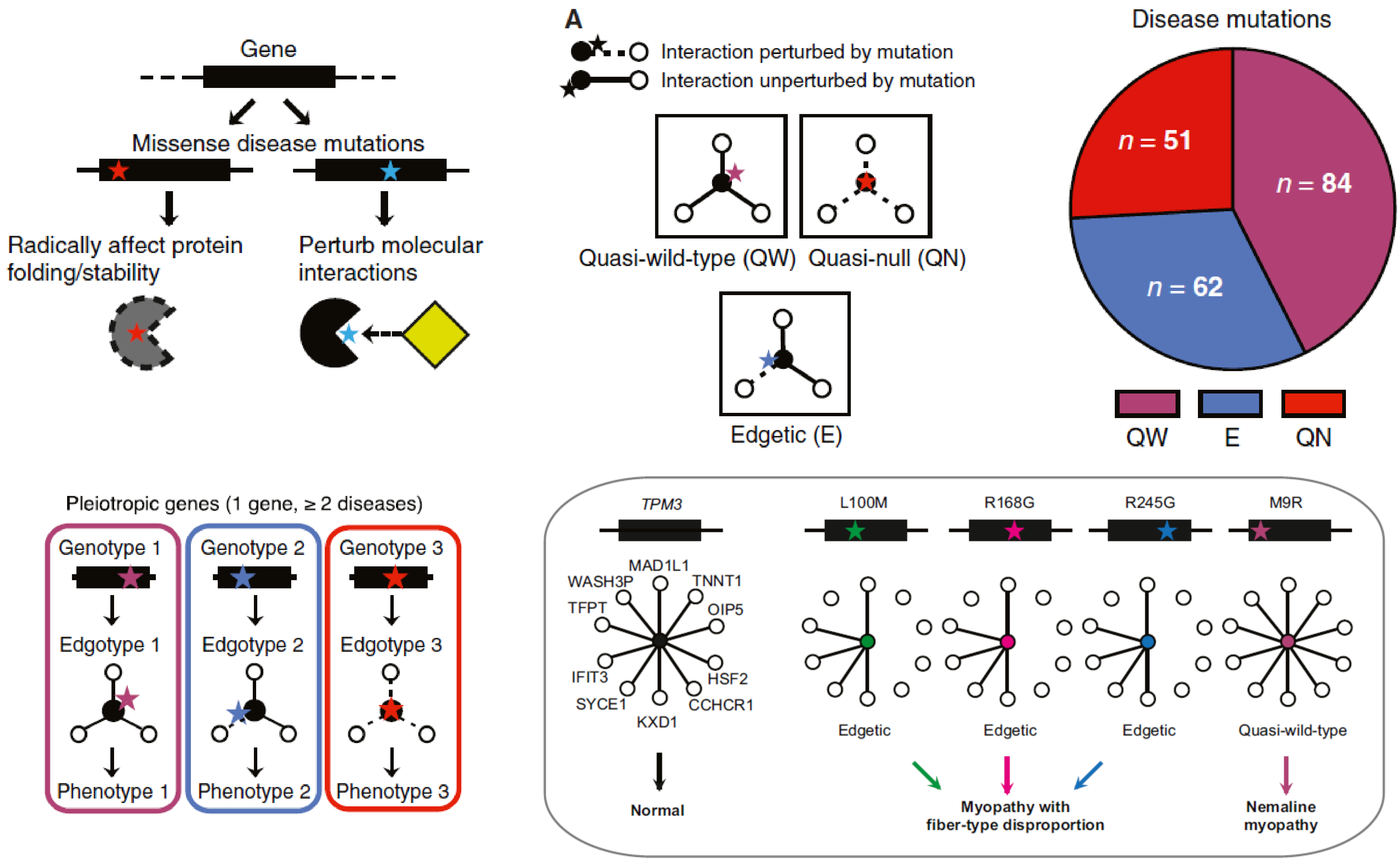
Missense disease mutations: stability or PPI?

Table 1 | **Human diseases caused by defects in protein folding, stability and aggregation**

Disease	Protein affected	Description	References
Cystic fibrosis	Cystic fibrosis transmembrane conductance regulator (CFTR)	The ΔPhe508 mutant has wild-type activity, but impaired folding in the endoplasmic reticulum leads to degradation.	97
α1 Antitrypsin deficiency	α1 Antitrypsin (also known as SERPINA1)	80% of Glu342Lys mutants misfold and are degraded. Pathology is due to aggregation in patients with a reduced degradation rate.	97
SCAD deficiency	Short-chain acyl-CoA dehydrogenase (SCAD)	Impaired folding of Arg22Trp mutants leads to rapid degradation.	98
Alzheimer disease	Presenilin, γ-secretase	Mutations cause incorrect cleavage by the γ-secretase protease to produce the amyloid β-peptide; this aggregates into extracellular amyloid plaques.	99,100
Parkinson disease	α-Synuclein	Oxidative damage causes misfolding and aggregation. Hereditary forms are linked to deficiency in ubiquitin-mediated degradation.	101
Huntington disease	Huntingtin	CAG expansions in the Huntingtin gene lead to an abundance of polyglutamine fragments that aggregate and associate non-specifically with other cellular proteins.	101,102
Sickle cell anaemia	Haemoglobin	The Glu6Val mutation leads to aggregation in red blood cells.	103

De Pristo (2005) *Nat Rev Genet*

Missense disease mutations: stability or PPI?



The effects of missense disease mutations on molecular interactions could range from no apparent detectable change in interactions (**quasi-WT**), to specific loss of some interactions (**edgetic**), to an apparent complete loss of interactions (**quasinull**)

Sahni (2015) *Cell*

Prediction of missense variant effect

Applications
- Disease gene discovery
- Clinical sequencing // ~11,000 nsSNVs per individual, including rare
- Evolutionary, population genetics
- Protein design

Missense effect is diverse; experiment is not feasible. **What experiment?**
In vivo:
- Clinical impact // rare, context-dependent, inheritance mode
- Model organisms // applicability?

In vitro:
- Functional assay // applicability?

In silico: Damaging | Tolerated, Benign
- Data sources and features
- Prediction methods
- Evaluation

Prediction of missense variant effect

Data sources

Clinical impact . . . . pathogenic

- ClinVar, HGMD

Biochemical assays. . . . . .functional

- Papers, Protein Mutant Database

Deep mutational scans. .functional

- Papers, MAVEdb

Population data. . . . . deleterious

- dbSNP, ExAC/gnomAD, other species

Phylogenetic data. . . . . deleterious

- NCBI nr, UniPto UCSC MultiZ

Prediction of missense variant effect

Features

---

1. Substitution

- Conservative / radical (BLOSUM, Grantham score)
- Volume, hydrophobicity change

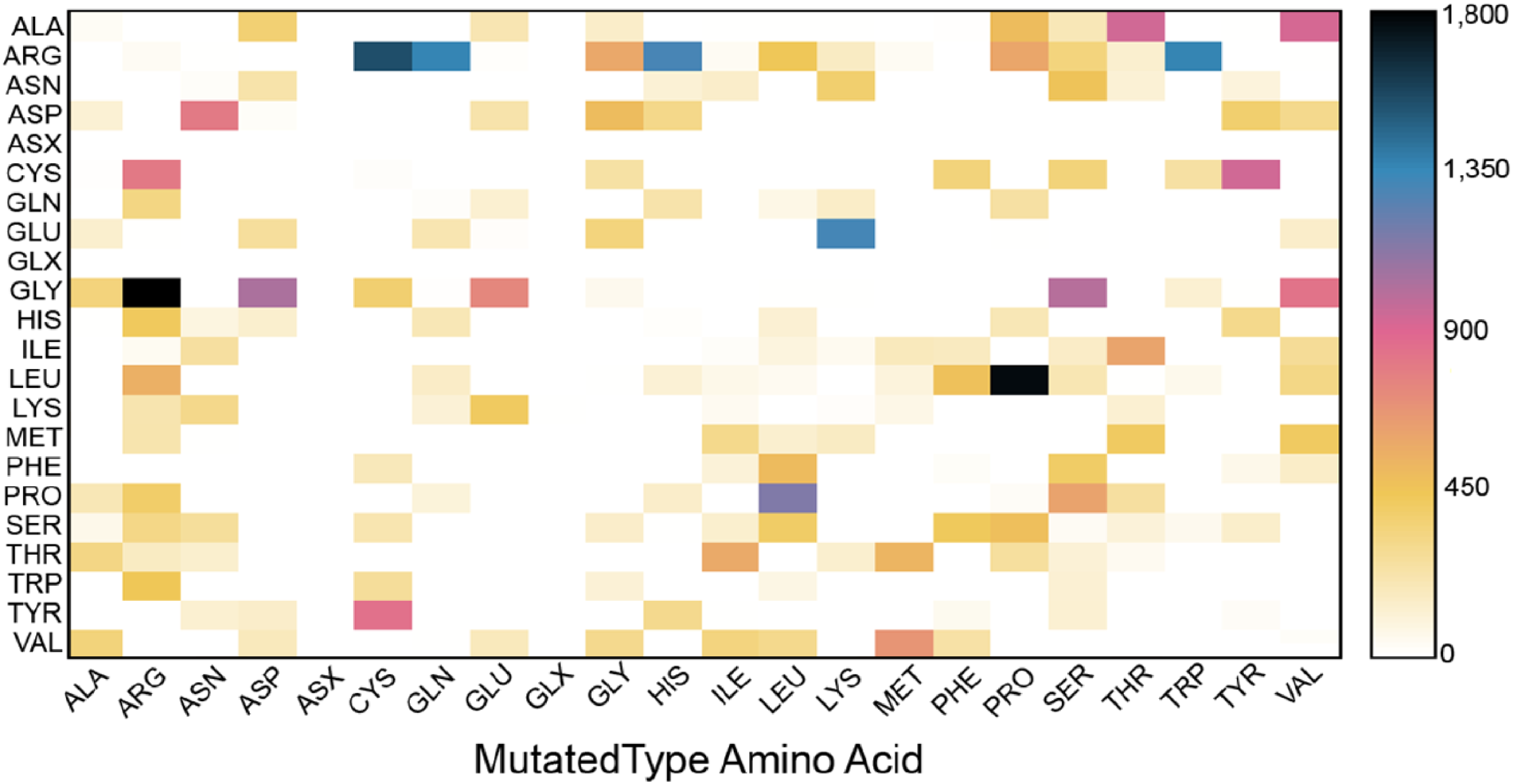
2. Site

- Conservation
- Location: core / surface (Relative Surface Area)
- Contacts: protein, ligand, DNA/RNA
- Secondary structure, disorder
- B-factor

3. Protein

- Number of interactions
- Number of PubMed references

Missense variants in human disease

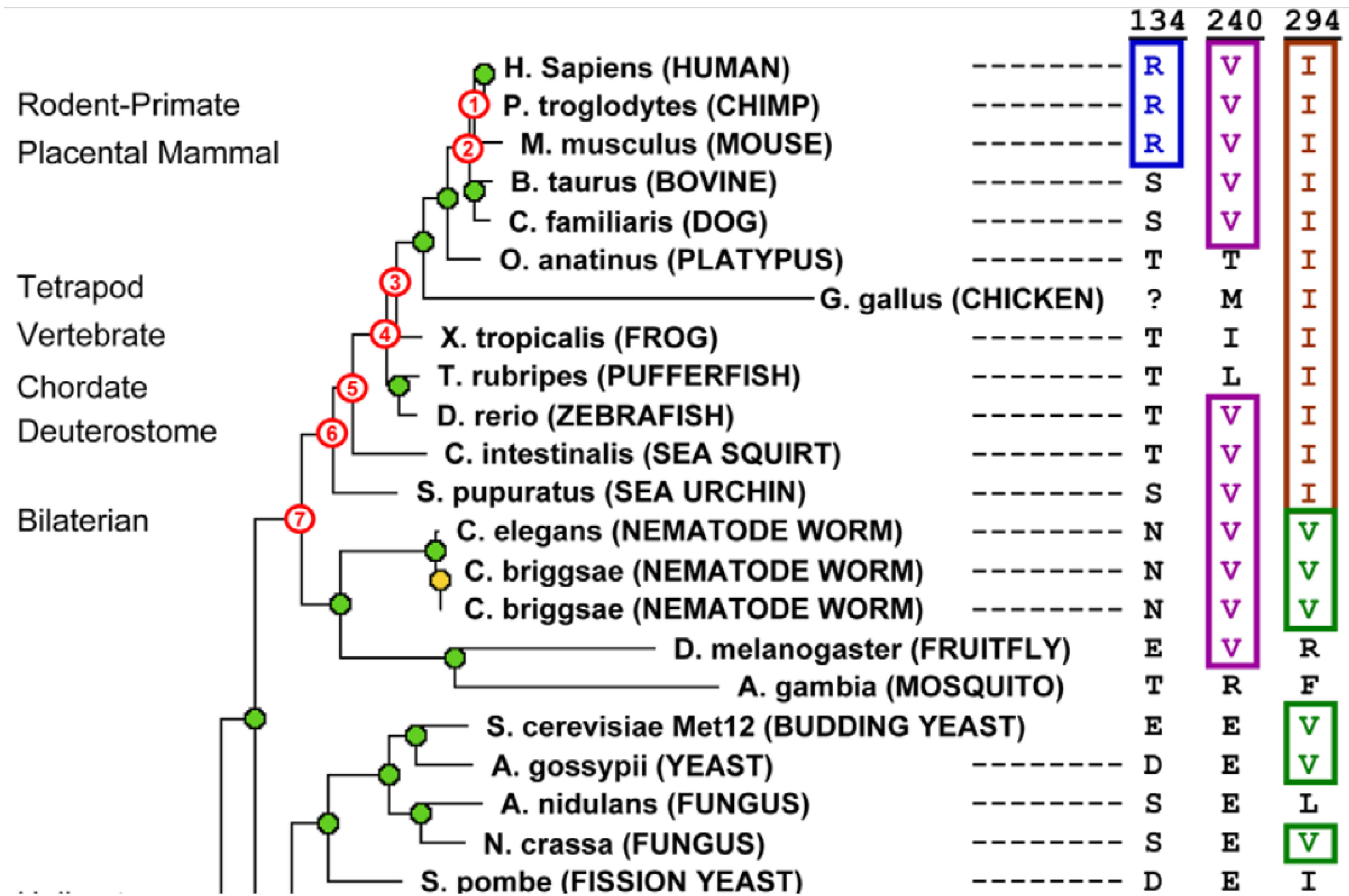


Exercise: list top 10 most frequent disease-causing missense variants

Peterson (2013) *J Mol Biol*

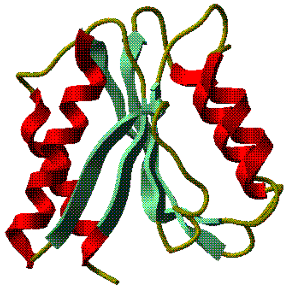
Prediction of missense variant effect

Multiple Sequence Alignment: evolutionary record



Marini (2010) *PLOS Genet*

Prediction of missense variant effect



Protein → Multiple Sequence Alignment

N	E	L	V	T	L	T	C	L	A	R	G	F	S	-	P	K	D	V	L	V	R	W	L
R	E	S	A	T	I	T	C	L	V	T	G	F	S	-	P	A	D	V	F	V	Q	W	M
G	G	S	L	R	L	S	C	V	A	S	G	I	T	-	F	S	G	Y	D	M	Q	W	V
T	P	G	L	T	L	T	C	T	V	S	G	F	S	-	L	S	S	Y	D	M	G	W	V
G	Q	K	A	K	M	R	C	I	P	E	-	-	-	-	K	G	H	P	V	V	F	W	Y
G	Q	E	A	T	L	W	C	E	P	I	-	-	-	-	S	G	H	S	A	V	F	W	Y
G	Q	Q	V	T	L	S	C	F	P	I	-	-	-	-	S	G	H	L	S	L	Y	W	Y
R	K	D	V	S	L	T	C	L	V	V	G	F	N	-	P	G	D	I	S	V	E	W	T
G	Q	K	L	T	L	K	C	Q	Q	N	-	-	-	-	F	N	H	D	T	M	Y	W	Y
R	D	K	A	T	F	T	C	F	V	V	G	S	D	-	L	K	D	A	H	L	T	W	E
S	K	S	A	T	L	T	C	R	V	S	N	M	V	N	A	D	G	L	E	V	S	W	W
G	A	R	T	S	L	N	C	T	F	S	D	-	-	-	S	A	S	Q	Y	F	W	W	Y
G	A	S	L	Q	L	R	C	K	Y	S	Y	-	-	-	S	A	T	P	Y	L	F	W	Y
N	G	A	P	K	L	T	C	L	V	V	D	L	E	S	E	K	N	V	N	V	T	W	N
E	A	T	V	T	L	T	C	V	V	S	N	-	-	A	P	Y	G	V	N	V	S	W	T

Profile

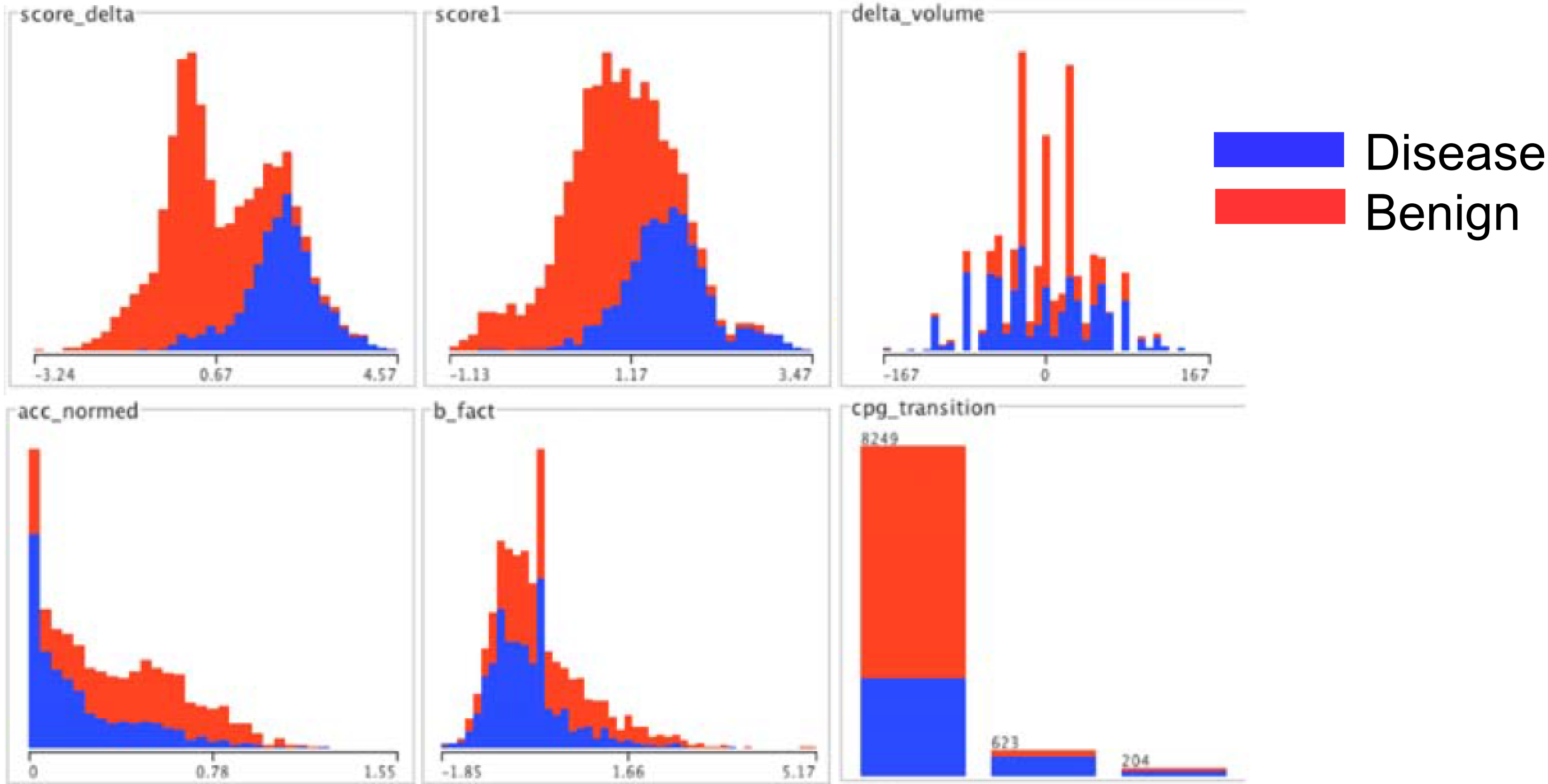
↓

Ala	-1.2	1.1	-0.6	-0.8	0.3	...	...
Arg	0.6	-0.3	-0.3	-0.5	0.6	...	...
Asn	-1.1	-0.5	-0.5	-0.7	0.4	...	...
Asp	-0.9	-0.3	-0.3	-0.5	0.6	...	...
Cys	0.4	-0.5	0.6	0.8	-0.3	...	...
Gln	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...

PSIC (Position Specific Independent Counts) profile scores matrix

Sunyaev (1999) *Protein Eng*

Prediction of missense variant effect

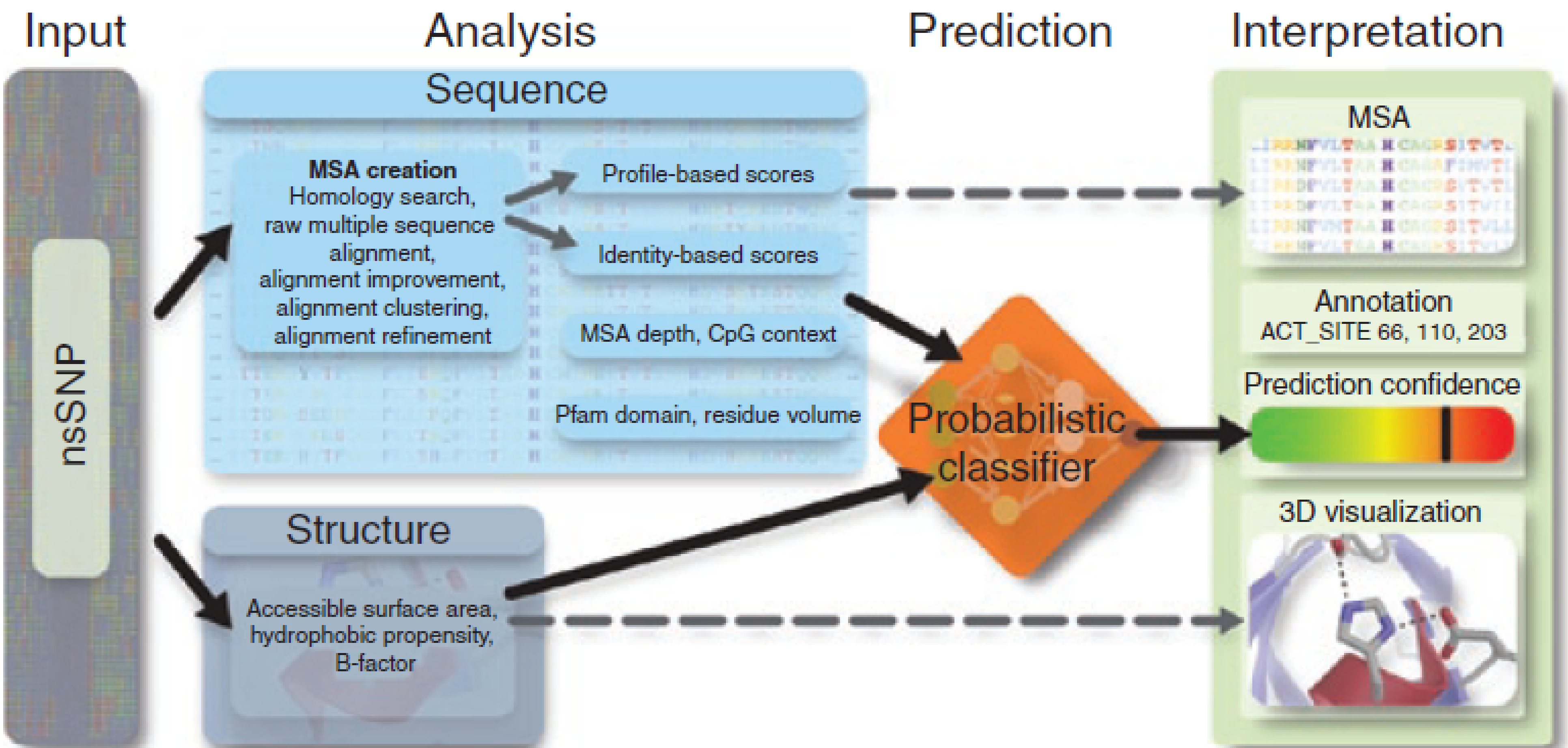


Examples of predictive features used by PolyPhen-2

score_delta: PSIC(AA1)-PSIC(AA2)
score1: PSIC(AA1)
delta_volume: change in side chain volume
cpg_transition: CpG context (0:no, 1: removes CpG, 2:creates)
acc_normed:* normalized accessible surface area // if 3D structure available
b_fact:* average temperature factor

Adzhubei (2010) *Nat Methods*

Prediction of missense variant effect

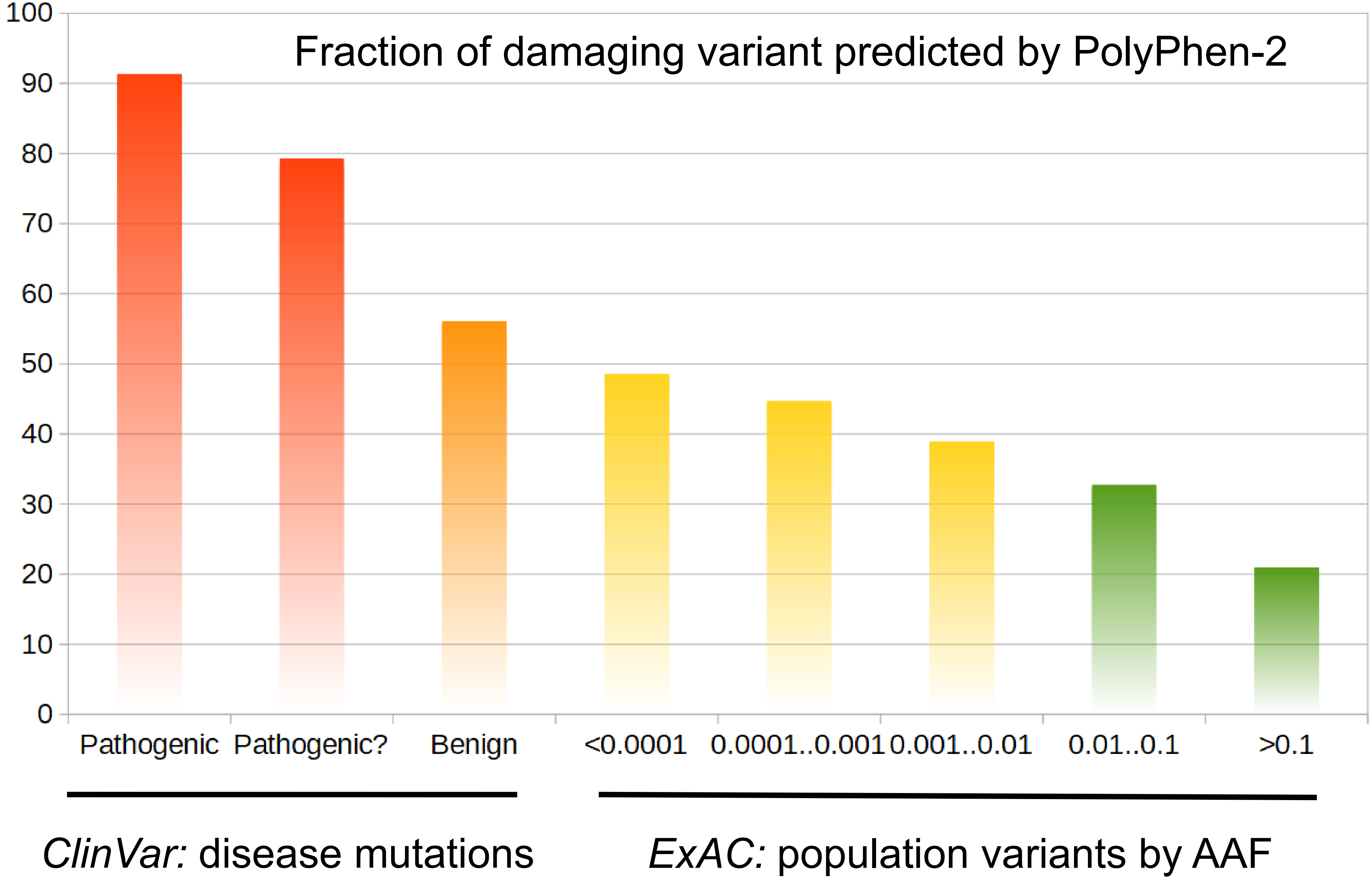


PolyPhen-2 prediction pipeline

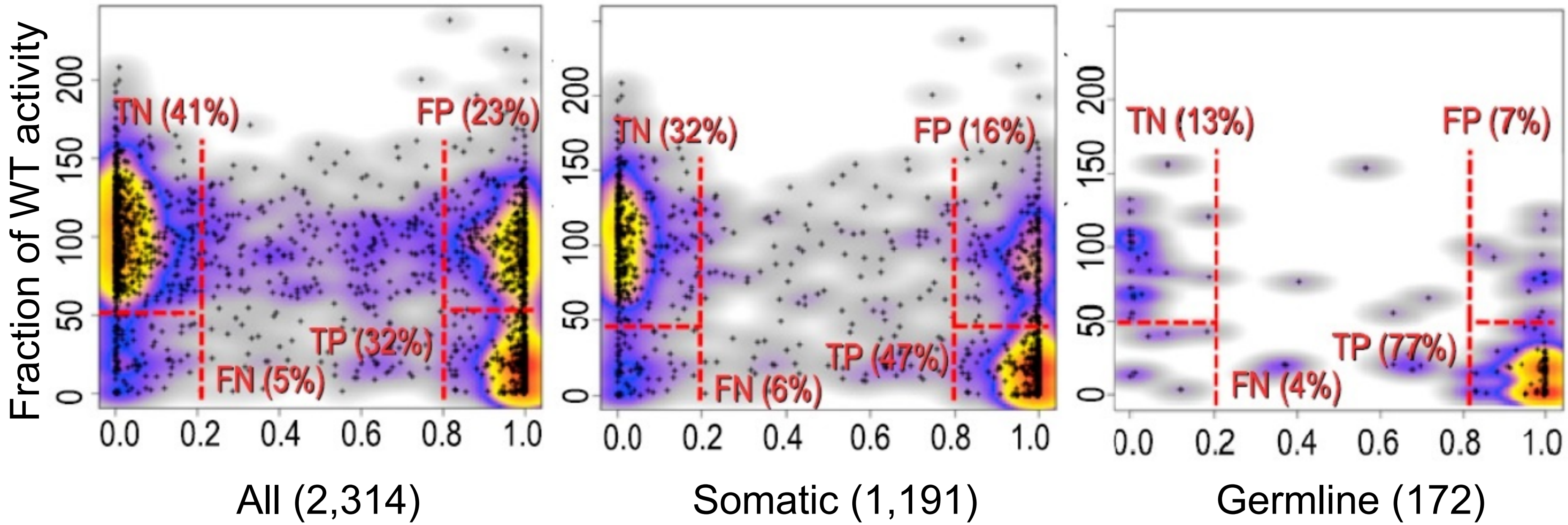
Training set (HumDiv): 3,155 disease mutations, 6,321 human-ortholog subst
Performance: FPR=10%, TPR=77%;   FPR=20%, TPR=92%

Adzhubei (2010) *Nat Methods*

Prediction of missense variant effect



Prediction of missense variant effect



<u>What do we predict?</u>

- Experiment: *in vitro* activity of TP53 compared with predictions by PolyPhen-2, threshold: 50% of WT activity
- Low false negative prediction rate, but
- 42% of mutations predicted by PolyPhen2 to be damaging had little measurable consequence for TP53-promoted transcription
- The predictions do not effectively differentiate between mutations that are immediately clinically relevant (ablate or markedly reduce function), and those that are nearly neutral (decrease the function of the corresponding protein by 10%)

Miosge (2015) *PNAS*

Prediction of missense variant effect

Damaging, deleterious, pathogenic, detrimental

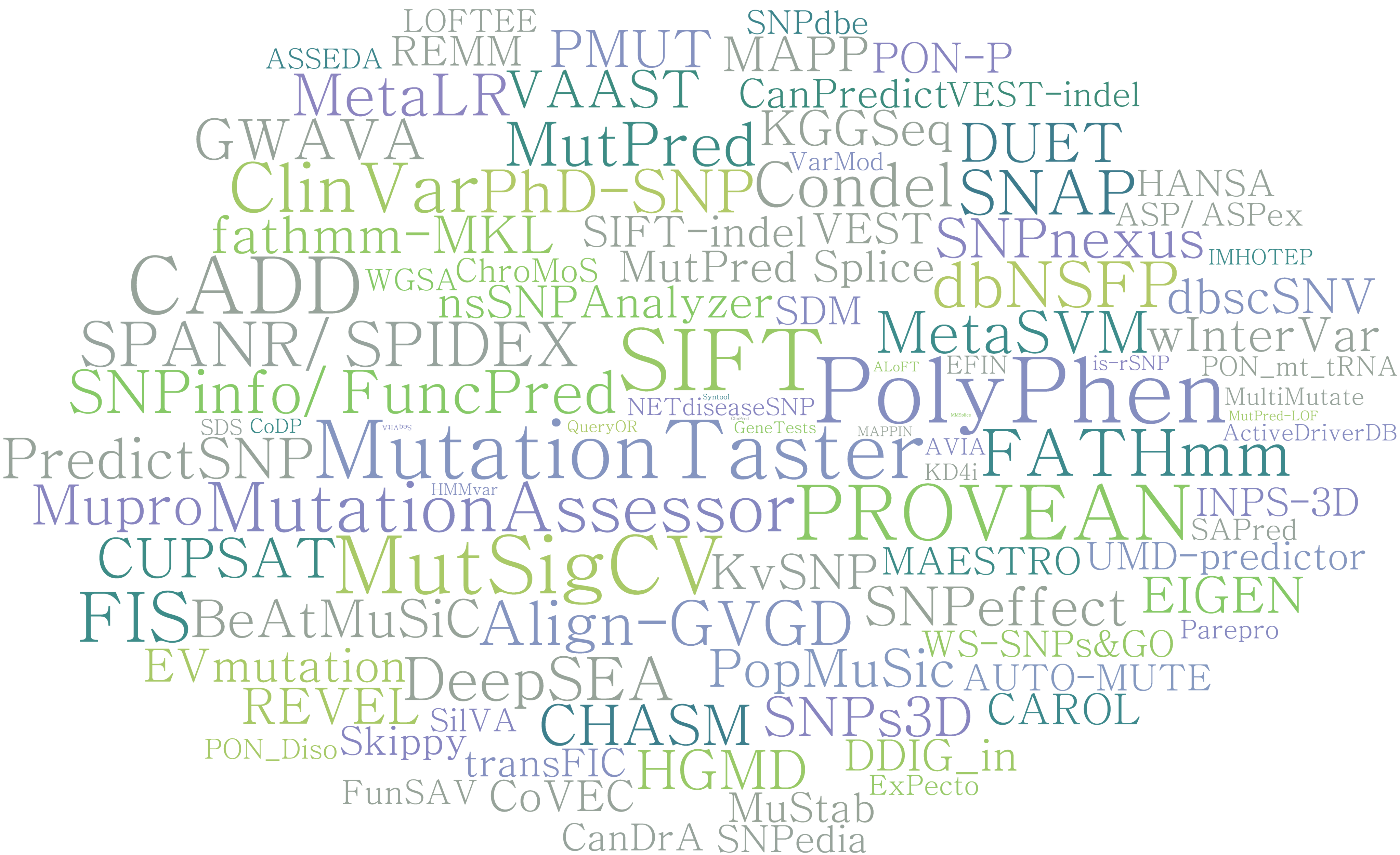
The effect of a missense mutation on an organism is always multifaceted and can be considered from multiple perspectives—**biochemical, medical, and evolutionary**. The relationship between the effects of amino acid substitution on protein activity, human health, and an individual's evolutionary fitness is not trivial.

A mutation that damages protein structure does not necessarily lead to a detectable human-disease phenotype, and a mutation that predisposes an individual toward a disease is not necessarily evolutionarily deleterious. <...> Substitutions leading to abnormal hemoglobin function that cause sickle-cell anemia are apparently negative from both biochemical and medical points of view. Nevertheless, they cannot be considered negative from an evolutionary point of view, because balancing selection has brought them to high frequency in many parts of the world as a result of malaria resistance in heterozygotes.

To clearly distinguish different aspects of negative mutations, we use the term **damaging** to refer to a mutation that decreases protein activity, the term **detrimental** to refer to a mutation that predisposes an individual toward a disease, and the term **deleterious** to refer to a mutation that has been subject to purifying selection.

Kryukov (2007) *Am J Hum Genet*

Prediction of missense variant effect



https://genomeinterpretation.org/vipdb

NEW!

Prediction of missense variant effect

- **Predictions for the whole proteome**: dbNSFP, 84 mln missense and splicing site SNVs
- **Ensemble (meta-) predictors**: MetaSVM, MetaLR, ReVel, M-CAP, etc
- **Neural networks and other ML techniques**: PrimateAI, ~380,000 common missense variants from humans and primates, gradient boosting tree classifier
- **Covariation**: EVmutation accounts for epistasis by explicitly modeling interactions between all the pairs of residues
- **Prediction of quantitative effect**: Envision 21,026 variant effect measurements from 9 large-scale experimental mutagenesis datasets
- **Clinical applicability**: M-CAP, 9 tools, 7 conservation scores, 298 features derived from MSA, gradient boosting tree classifier

Prediction of missense variant effect

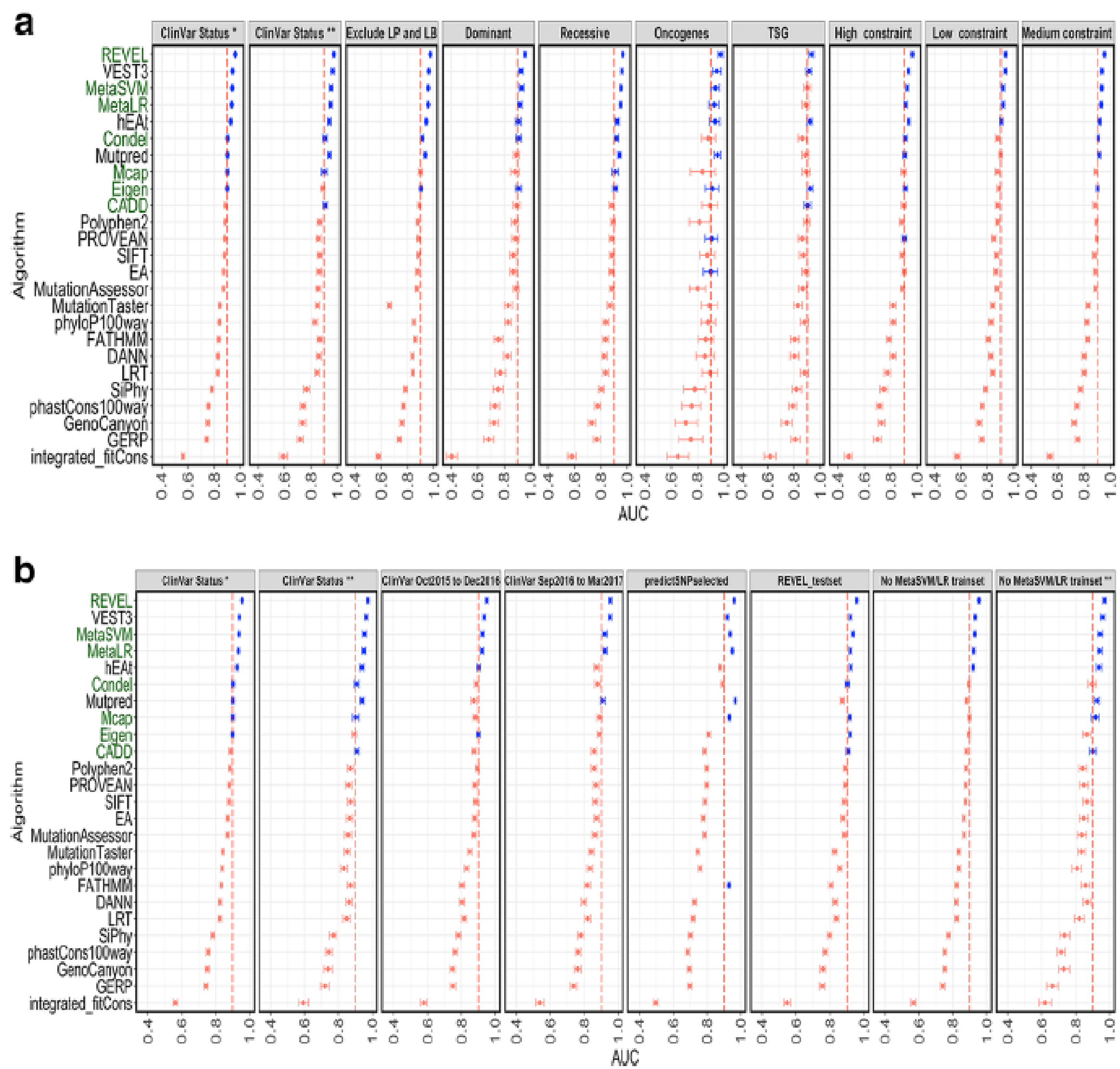
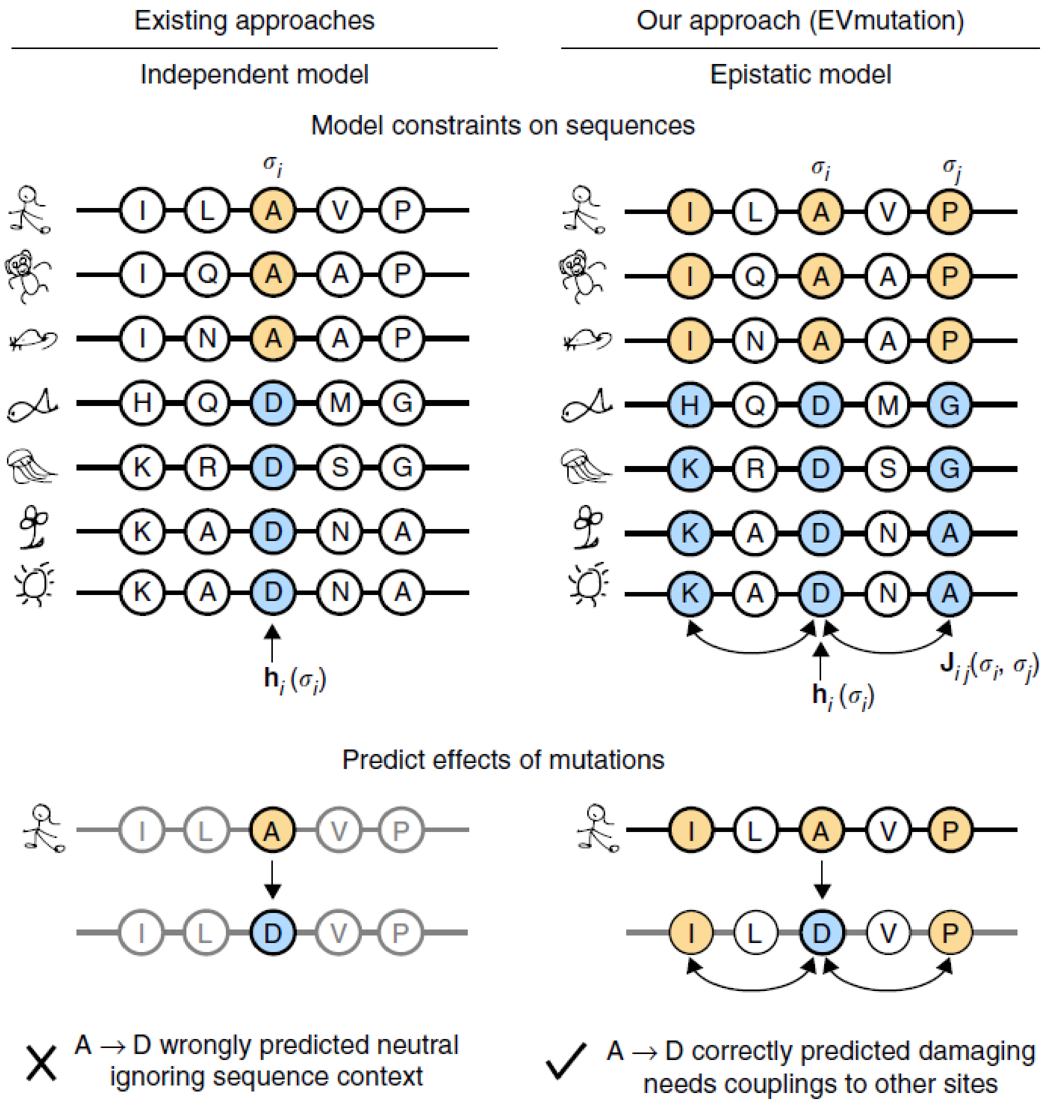


Fig. 3 Performance analysis of algorithms. The AUC of a ROC are plotted for 25 algorithms. *Vertical dotted line* indicates an AUC of 0.9 and 99% confidence intervals for each AUC are shown. *Blue dots* indicate AUC > 0.89. **a** AUCs of the algorithms across different datasets shown in the panels and described in text. **b** AUCs of the algorithms across different datasets (represented in panels) to address type I circularity as described in text. The same plots for ClinVar Status * and ClinVar Status ** as in Fig. 3a are used in 3b for comparison. Any instance of ** represents variants with ClinVar review status of two stars or above. Ensemble predictors are indicated by *dark green labels* on the *y-axis*

Prediction of missense variant effect



Inferring context-dependent effects of mutations from sequences. Evolution has generated diverse families of proteins and RNAs with varied sequences that perform a common function. An unsupervised probabilistic model trained to generate the natural diversity in a multiple sequence alignment of a family can be used to predict the relative favorability of unseen mutations. Existing models describe functional constraints on each position i in a sequence σ independently, averaging over the effect of background positions j. This can lead to incorrect predictions of neutrality. Our approach infers a global probability model with pairwise interactions between positions i and j (J_{ij}) as well as background biases at single positions (h_i).

Hopf (2017) *Nat Biotech*

Prediction of inframe indels effect

MS2 COAT PROTEIN

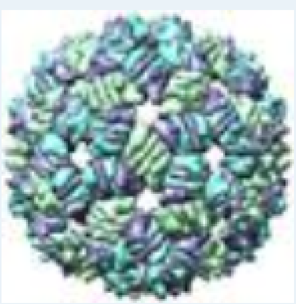
Query:

PDB ID: **2BU1**

Chain ID: A

EC number:

BACTERIOPHAGE FR CAPSID



Subject:

PDB ID: **1FR5**

Chain ID: A

EC number:


```
2BU1.A   61 KVEVPKVATQTVGGVELPVAAWRSYLNMELTIPIFATNSDCELIVKAMQGLLKDGNPIPS 120
            |||||||||    ||||||||||||.|||||||.|||| || |||||.||  | ||||  .
1FR5.A   61 KVEVPKVAT----GVELPVAAWRSYMNMELTIPVFATNDDCALIVKALQGTFKTGNPIAT 116
```

Prediction of inframe indels effect

	Insertions, duplications	Deletions
ClinVar, 21 Oct 2019 (hg38)		
Pathogenic, Likely pathogenic	303	1,193
Benign, Likely benign	306	483
Other	1,291	3,566
GnomAD 2.1.1 (hg38)		
AF_POPMAX<1%	30,489	79,023
AF_POPMAX≥1%	742	1,517
Unknown	7,389	10,640
Individual exome (GiaB)	228	275

Q: what is the most "famous" disease-causing inframe indel?

Prediction of inframe indels effect

Gene	*ClinVar*	*gnomAD*
KCNH2 Potassium Voltage-Gated Channel Subfamily H Member 2	Pathogenic (4) Unknown (8)	Rare (11)
PHOX2B Paired Like Homeobox 2B	Benign (7) Pathogenic (4) Unknown (2)	Common (2) Rare/Unknown (14)
CACNA1A Calcium Voltage-Gated Channel Subunit Alpha1 A	Benign (5) Pathogenic (2)	Common (4) Rare/Unknown (42)
FOXC1 Forkhead Box C1	Benign (5) Pathogenic (3) Unknown (4)	Common (2) Rare/Unknown (49)

Prediction of inframe indels effect

Method	Genome version	Coordinates	Implemen -tation	Publi- cation	Last update
VEST-Indel	37, 38	Genome	Web / Local	2016	2019
CADD	37, 38	Genome	Web / Local	2013	2019
SIFT Indel	37, 38	Genome	Web / Local	2013	2016
MutPred-Indel	37 ?	Protein	Web / Local	2019	-
DDIG-in	37	Genome	Web	2013	2017
PROVEAN	37	Genome	Web / Local	2012	2015

Prediction of inframe indels effect

Method	ML	Best features
VEST-Indel	Random forest	Log10 of count of publications in PubMed where gene name is mentioned, Exon Conservation, protein local regional sequence composition
CADD	SVM	cDNApos, ProtPos, PolyPhenVal, SIFTVal, Relative position in coding sequence
SIFT Indel	Decision tree	Repeat, DNA Conservation score, Protein disorder region, Fraction of all Pfam domains affected due to indel
MutPred-Indel	Neural Network	PSSM*, sequence conservation indices, number of homologs in the human and mouse genomes, relative position in protein
DDIG-in	SVM	Disorder, ASA*, DNA Conservation, Neff*, Probabylity of sheet
PROVEAN	Not ML	PROVEAN score

* PSSM - position-specific scoring matrix, ASA - solvent accessible surface area, Neff - number of effective homologous sequences aligned to residues

Prediction of inframe indels effect

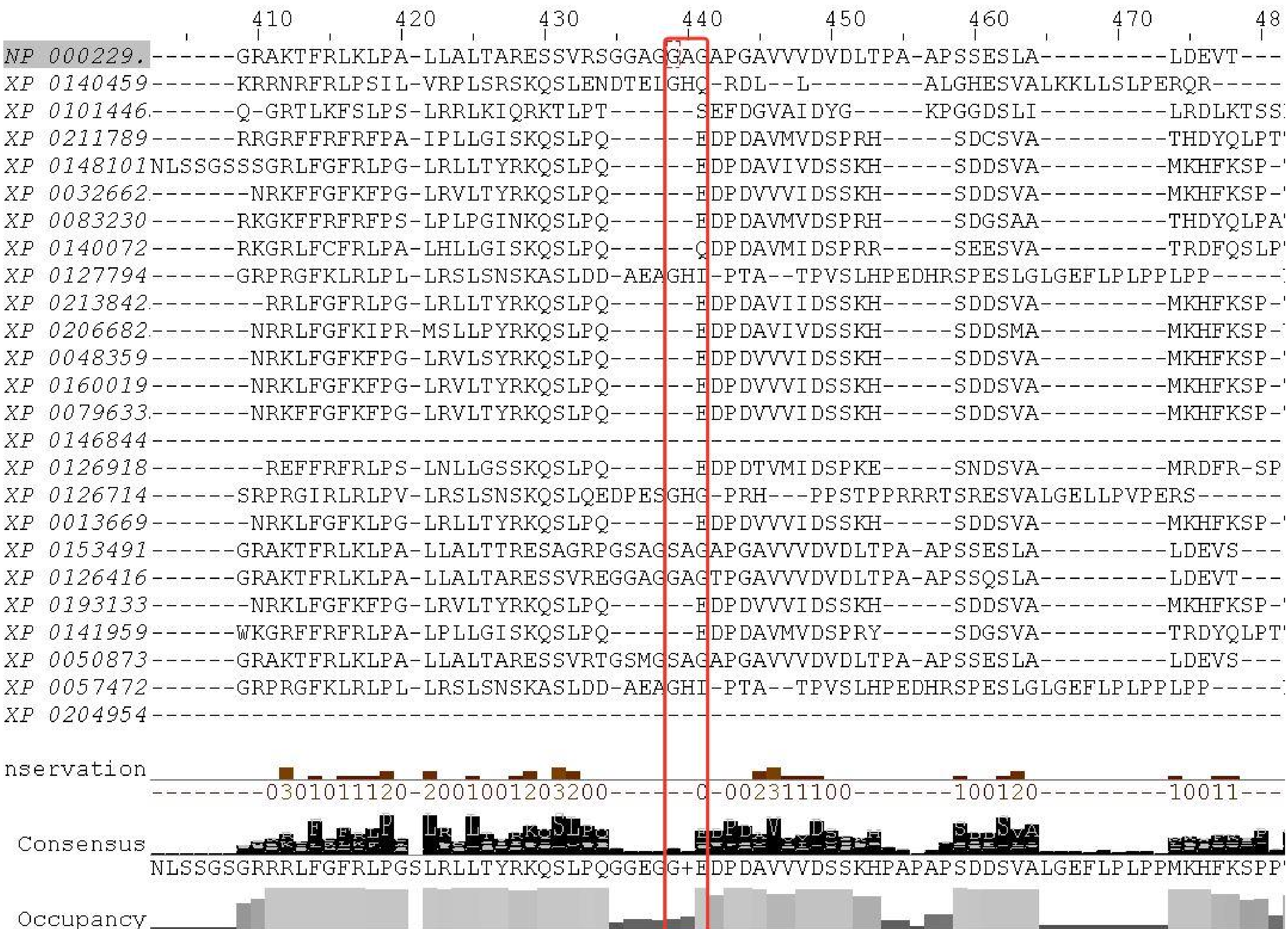
Meta-Predictors that Combine Classifications of Multiple Methods

In these Boolean expressions, each method is represented by a variable X_i, which is set to TRUE when the method classifies an example as pathogenic and FALSE when the method classifies an example as benign. For combinations of two methods, candidate meta-predictors were (X_1 and X_2) and (X_1 or X_2). For combinations of three methods, candidate meta-predictors (X_1 and X_2 and X_3), (X_1 or X_2 or X_3),(X_1 or X_2 or X_3), ((X_1 and X_2) or X_3), ((X_1 or X_2) and X_3), ((X_1 and X_3) or X_2),((X_1 or X_3) and X_2), ((X_2 and X_3) or X_1), ((X_2 or X_3) and X_1). For combinations of four methods, there are 64 possible combinations (Supp. Table S4). We used a brute-force approach and limited the number of methods in the meta-predictor to a maximum of four to avoid a combinatorial explosion. All possible four-way combinations of the five methods were explored.

Method	Sensitivity	Specificity	Balanced Accuracy
(VEST-indel AND PROVEAN) OR (CADD AND DDIG-in)	0.930	0.974	0.952
(VEST-indel OR CADD) AND PROVEAN	0.947	0.955	0.951
(VEST-indel OR CADD) AND (PROVEAN OR DDIG-in)	0.947	0.949	0.948
VEST-indel OR (CADD AND PROVEAN AND DDIG-in	0.930	0.955	0.942
VEST-indel OR (CADD AND DDIG-in)	0.930	0.949	0.939
VEST-indel OR (DDIG-in AND CADD)	0.930	0.949	0.939
VEST-indel OR (CADD AND PROVEAN)	0.947	0.929	0.938
(VEST-indel OR DDIG-in) AND PROVEAN	0.930	0.942	0.936

Douville (2016) *Hum Mutation*

Prediction of inframe indels effect

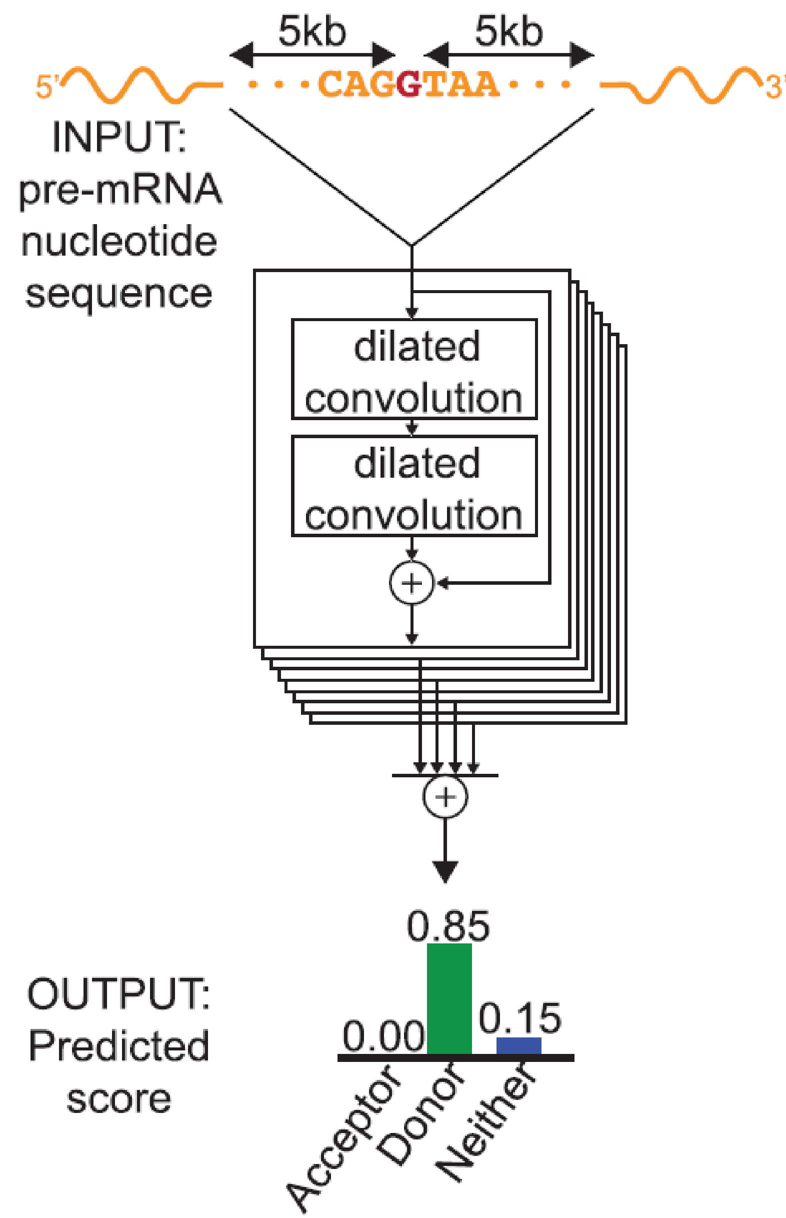


SpliceAI: predicting splicing from sequence

Essential splice variants disrupt canonical splice sites (GT, AG)
Cryptic splice variants: noncoding (intronic, synonymous) variants *outside* the canonical splice sites that disrupt the normal pattern of mRNA splicing

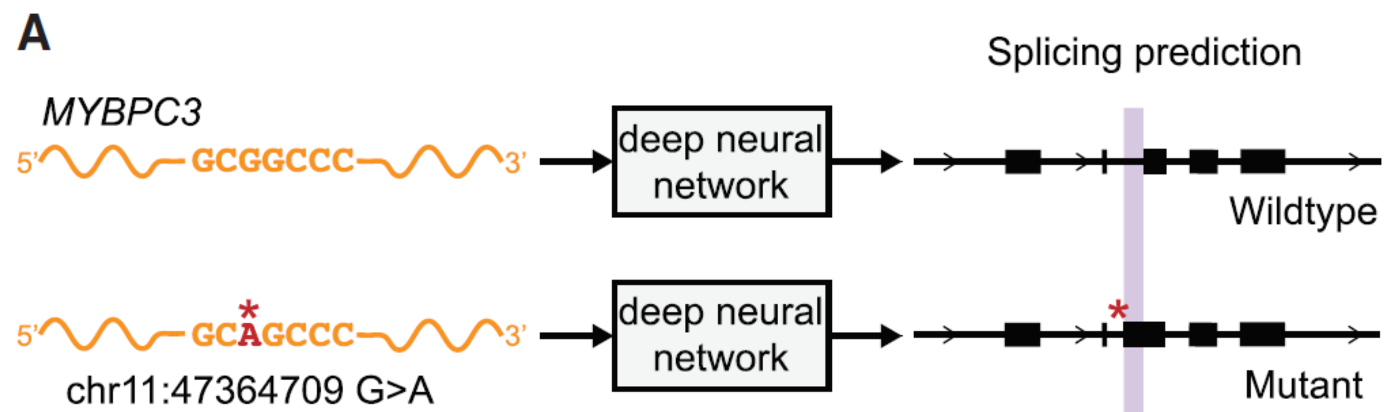
SpliceAI: a 32-layer deep neural network that accurately predicts splice junctions from an arbitrary pre-mRNA transcript sequence

Training set: pre-mRNA transcripts; algorithm learns the context of actual splicing sites



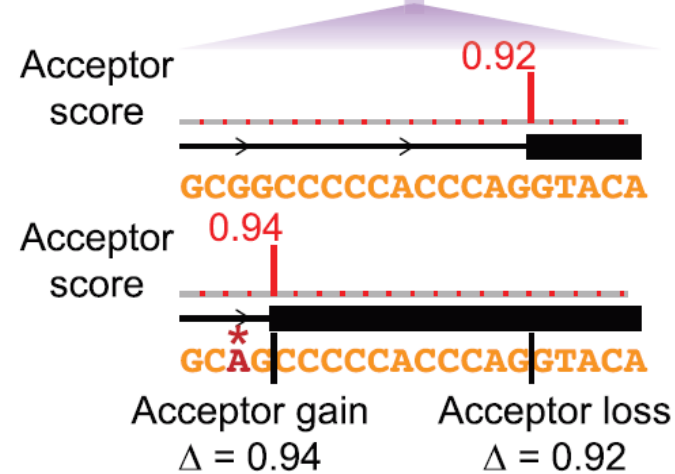
Jaganathan (2019) *Cell*

SpliceAI: predicting splicing from sequence



NM_000256.3(*MYBPC3*):c.1227-13G>A

Interpreted condition	Interpretation	Number of submissions
Hypertrophic cardiomyopathy	Pathogenic	2



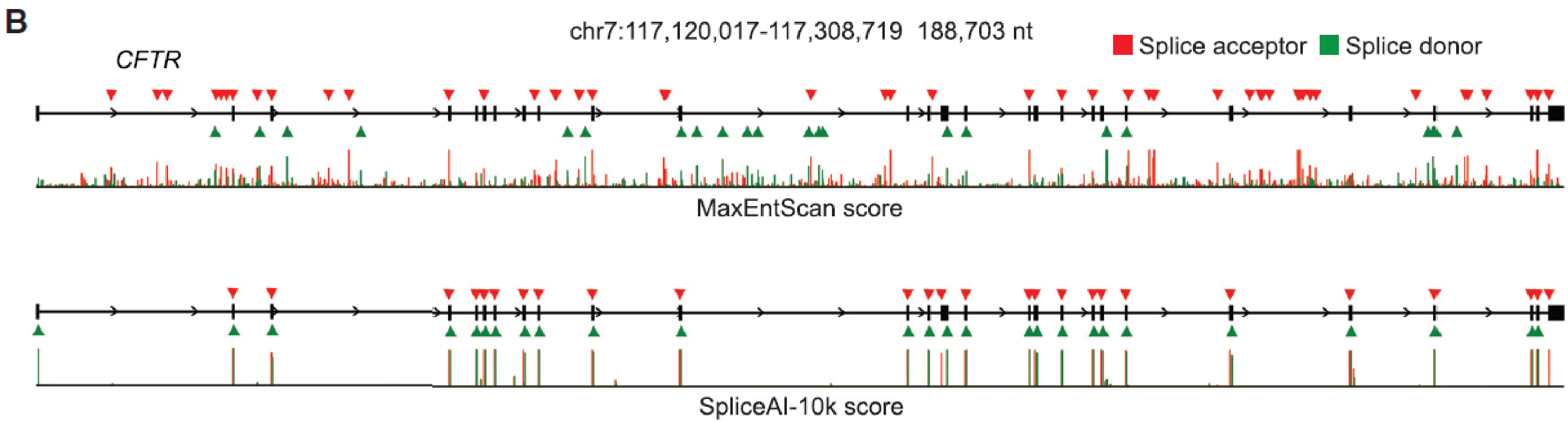
Input: position + flanks up to 5kbp

Output: P(acceptor), P(donor), P(neither)

SpliceAI-10k predicts acceptor and donor scores at each position in the pre-mRNA sequence of the gene with and without the mutation, as shown here for rs397515893, a pathogenic cryptic splice variant in the MYBPC3 intron associated with cardiomyopathy. The D score value for the mutation is the largest change in splice prediction scores within 50 nt from the variant.

Jaganathan (2019) *Cell*

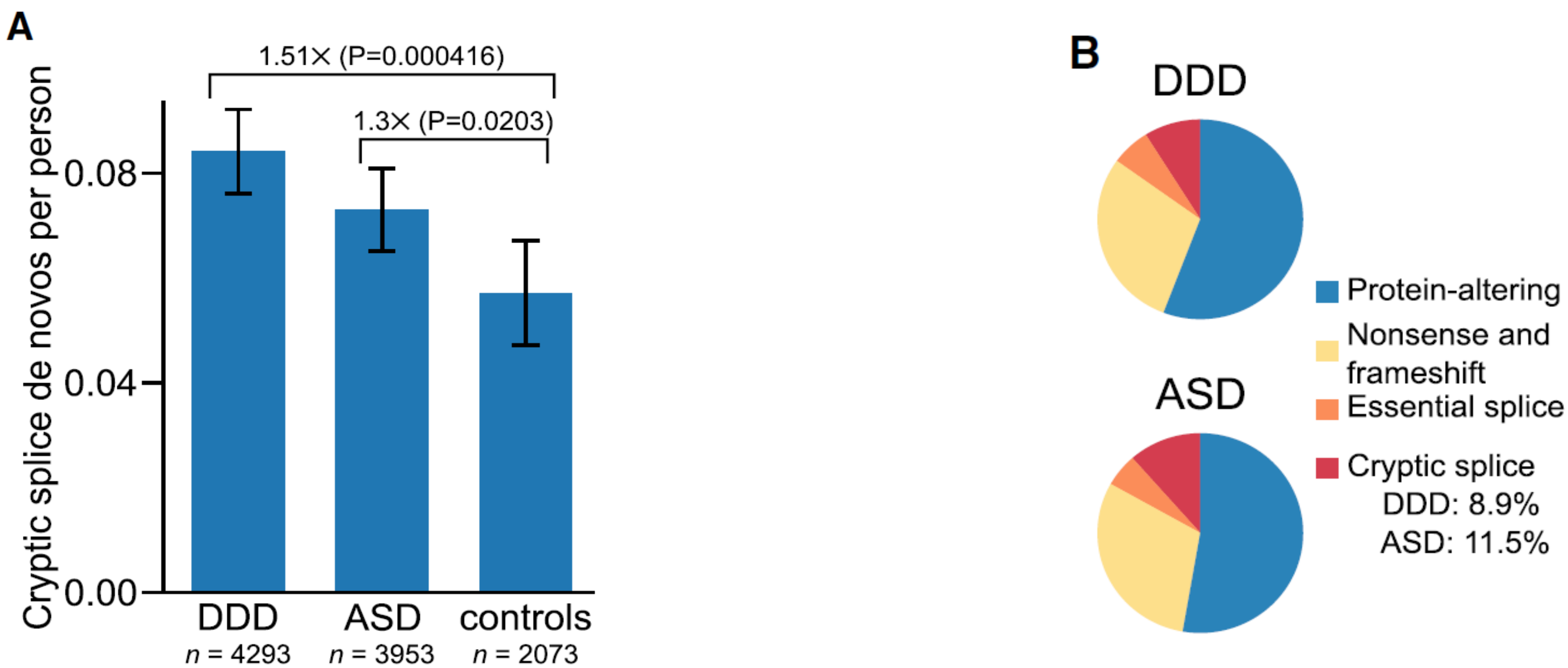
SpliceAI: predicting splicing from sequence



The full pre-mRNA transcript for the *CFTR* gene scored using MaxEntScan (top) and SpliceAI-10k (bottom) is shown, along with predicted acceptor (red arrows) and donor (green arrows) sites and the actual positions of the exons (black boxes). For each method, we applied the threshold that made the number of predicted sites equal to the total number of actual sites.

Jaganathan (2019) *Cell*

SpliceAI: predicting splicing from sequence



(A) Predicted cryptic splice de novo mutations per person for patients from the Deciphering Developmental Disorders cohort (DDD), individuals with autism spectrum disorders (ASDs) from the Simons Simplex Collection and the Autism Sequencing Consortium, as well as healthy controls.

(B) Estimated proportion of pathogenic de novo mutations by functional category for the DDD and ASD cohorts, based on comparison to controls.

Cryptic splicing may yield up to 10% of pathogenic variants in neurodevelopmental disorders

Jaganathan (2019) *Cell*

Regulatory elements in the human genome

Promoter: region (100-1000 bp) at the 5' end of genes where transcription factors and RNA polymerase bind to initiate transcription.

- Proximal promoters typically contain a CpG island
- Methylation of CpG islands silences genes

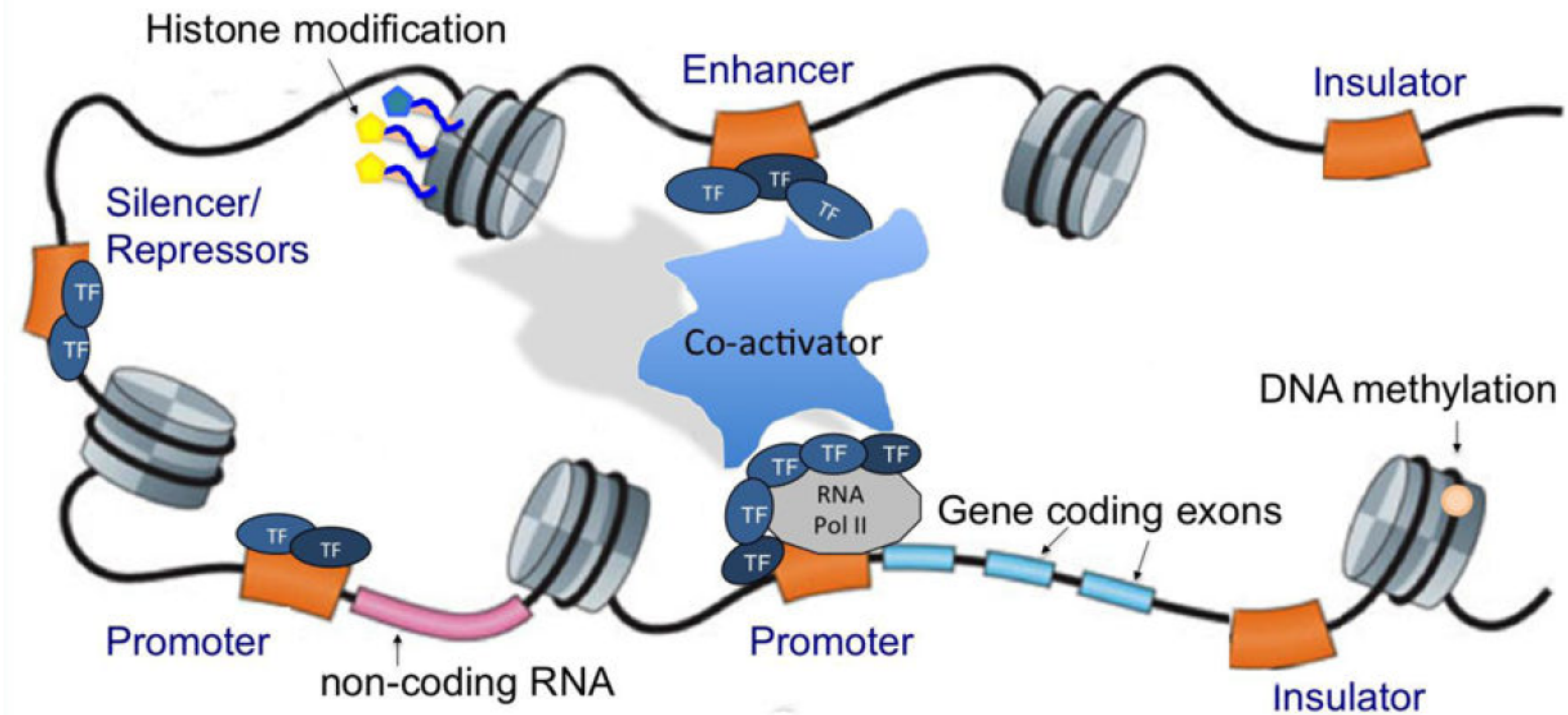
Enhancer: region (50-1500 bp) that binds transcription factors and interact with promoters to stimulate transcription of distant genes (<1Mbp)

- ~10^5 in the human genome (Penacchio 2013 *Nat Rev Genet*)
- Tissue-, time- or cell-specific
- Highly variable location (e.g., intron of an other distant gene)

Transcription factor binding motif/site: short genomic sequence that is known to bind to a particular transcription factor

- 1000-2000 TFs in the human genome
- 400-800 TFBS models (HOCOMOCO v.11)

Regulatory elements in the human genome

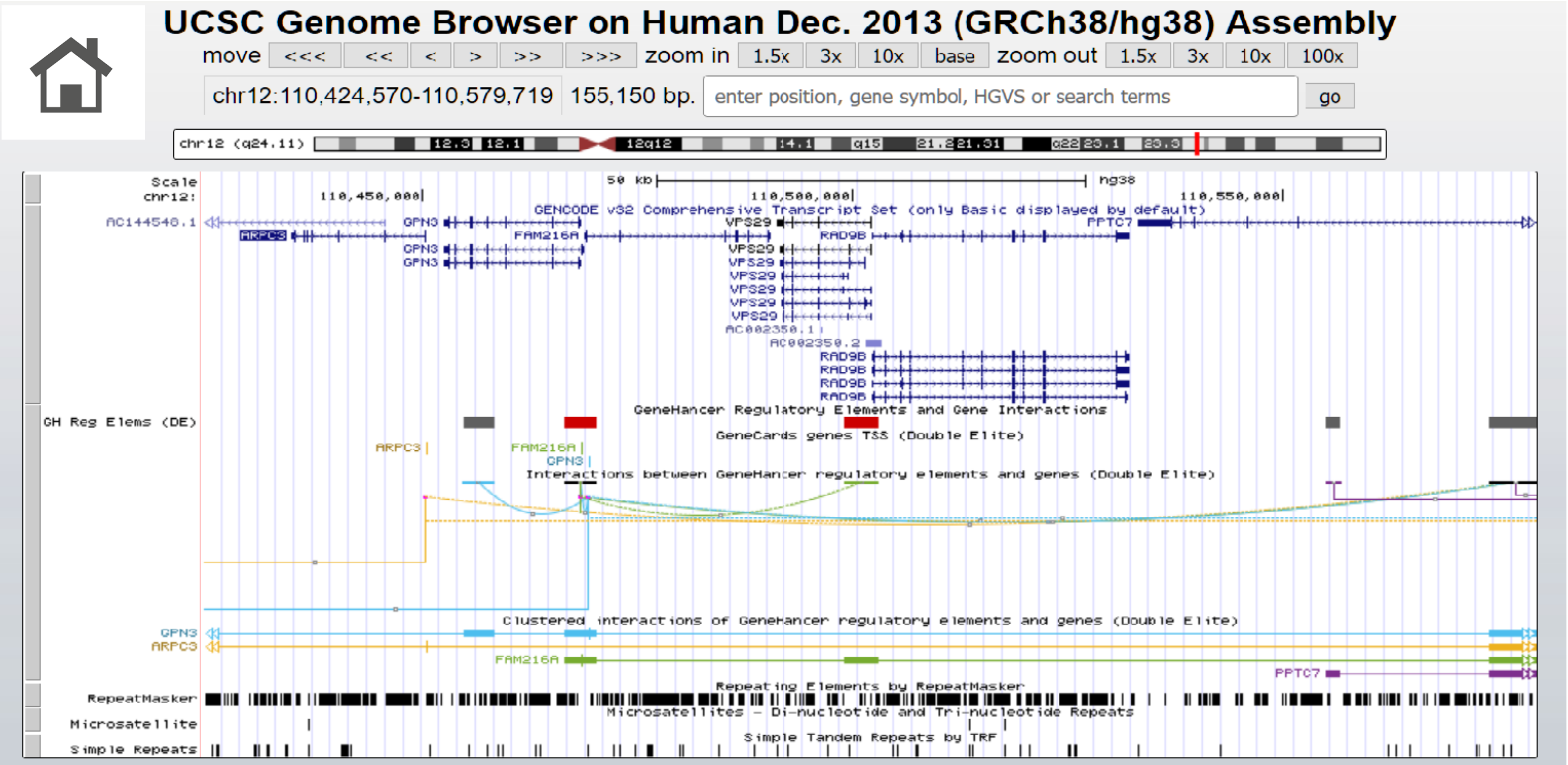


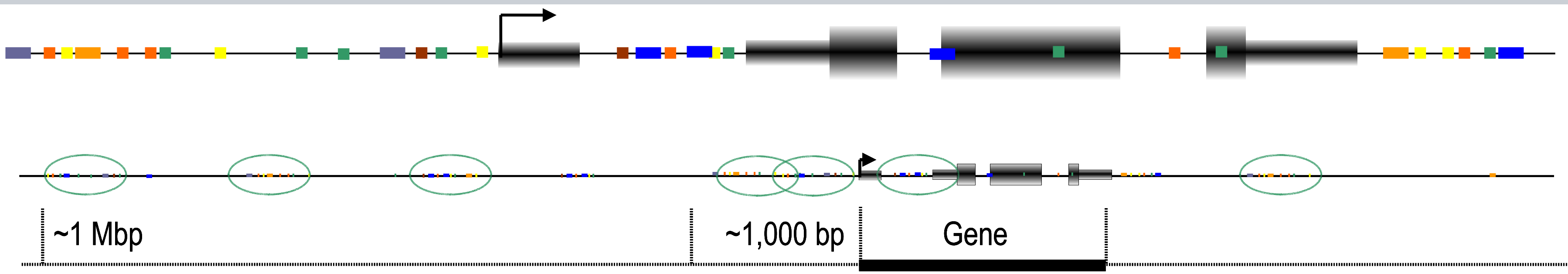
<u>Cis-regulatory elements:</u> **promoters** (100–1000bp) initiate the transcription of a target gene and are located immediately upstream of transcription start sites.

<u>Distal DNA regulatory elements</u>: Enhancers (50–1500bp), silencers, and insulators are DNA regulatory sequences, where transcription factors can bind and regulate expression rates of target genes. A complex of transcription factor and co-activators, mediated by **enhancers**, induce a conformational change of the chromatin structure, allowing the rapid production of specific genes depending on tissue/ cell-type and development-specific contexts. This lies in contrast to co-repressors, which serve to reduce gene expression by attaching to **silencers**. **Insulators** (300–2000bp) establish boundaries of gene expression by mediating loop formation and nucleosome modifications and thus prevent unneeded interactions of both enhancers and silencers with promoters

Lee (2018) *Hum Genet*

Regulatory elements in the human genome





Examples of non-coding functional variants

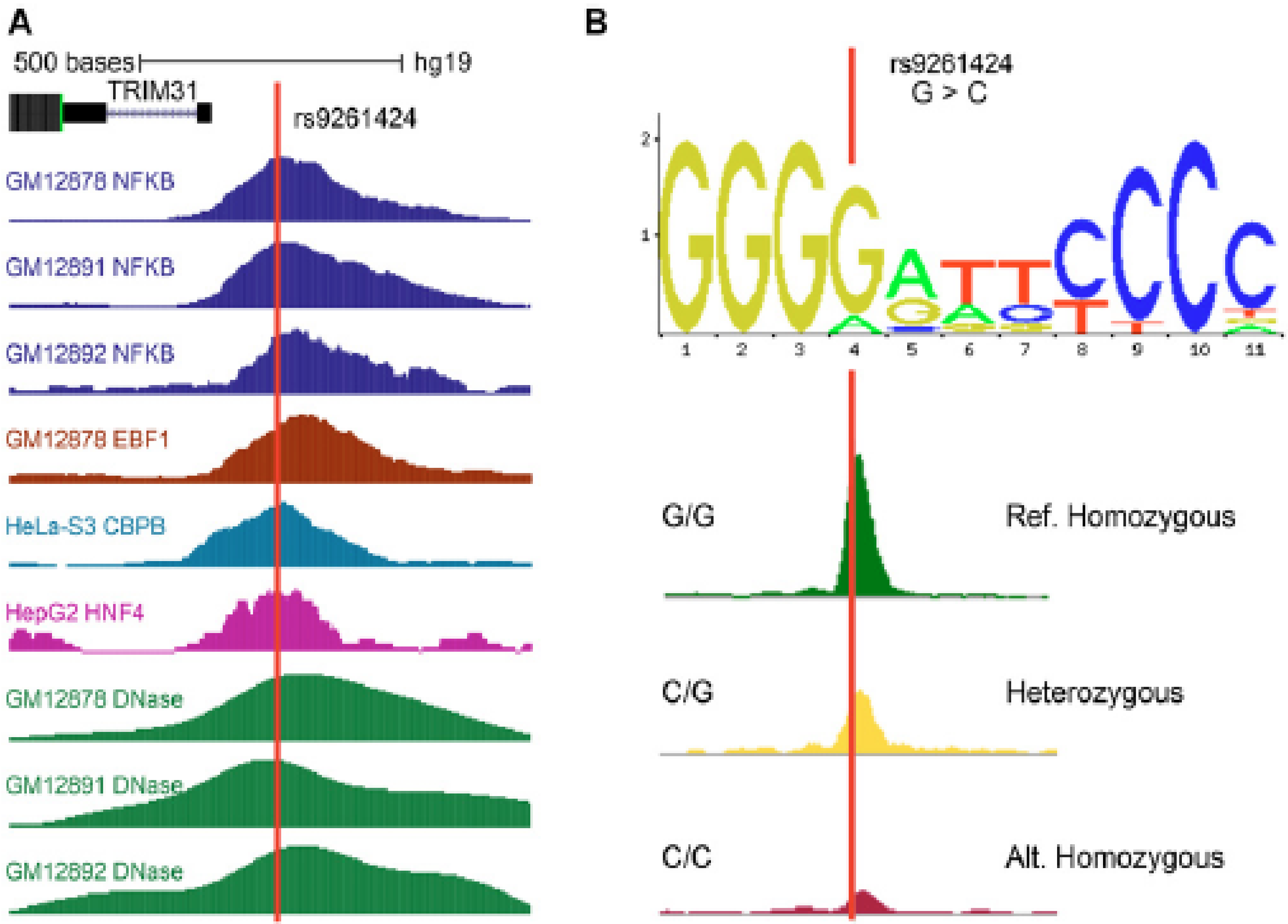
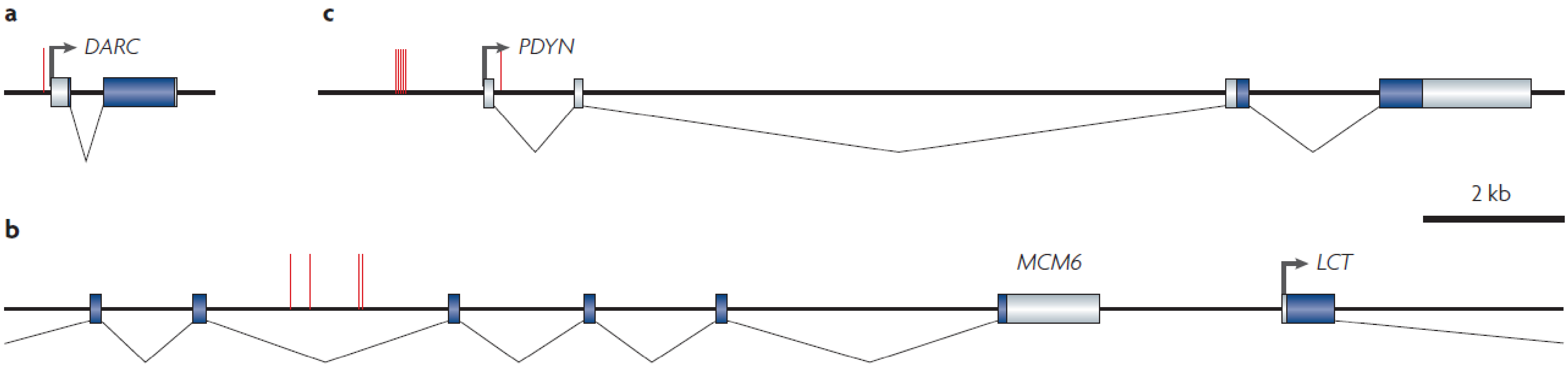


Figure 1. A SNV (rs9261424) overlapping many regulatory features. (*A*) This SNV falls within peak regions for many ChIP-seq factors as well as DNase-seq peaks from multiple cell lines. (*B*) The same SNV overlaps a motif match to the NFKB motif and has been shown to alter binding. The signal tracks represent ChIP-seq peaks of NFKB at the SNV site for three individuals: homozygous to reference allele (*G*), heterozygous, and homozygous to alternate allele (*C*) (Kasowski et al. 2010).

Boyle (2012) *Genome Res*

Examples of non-coding functional variants



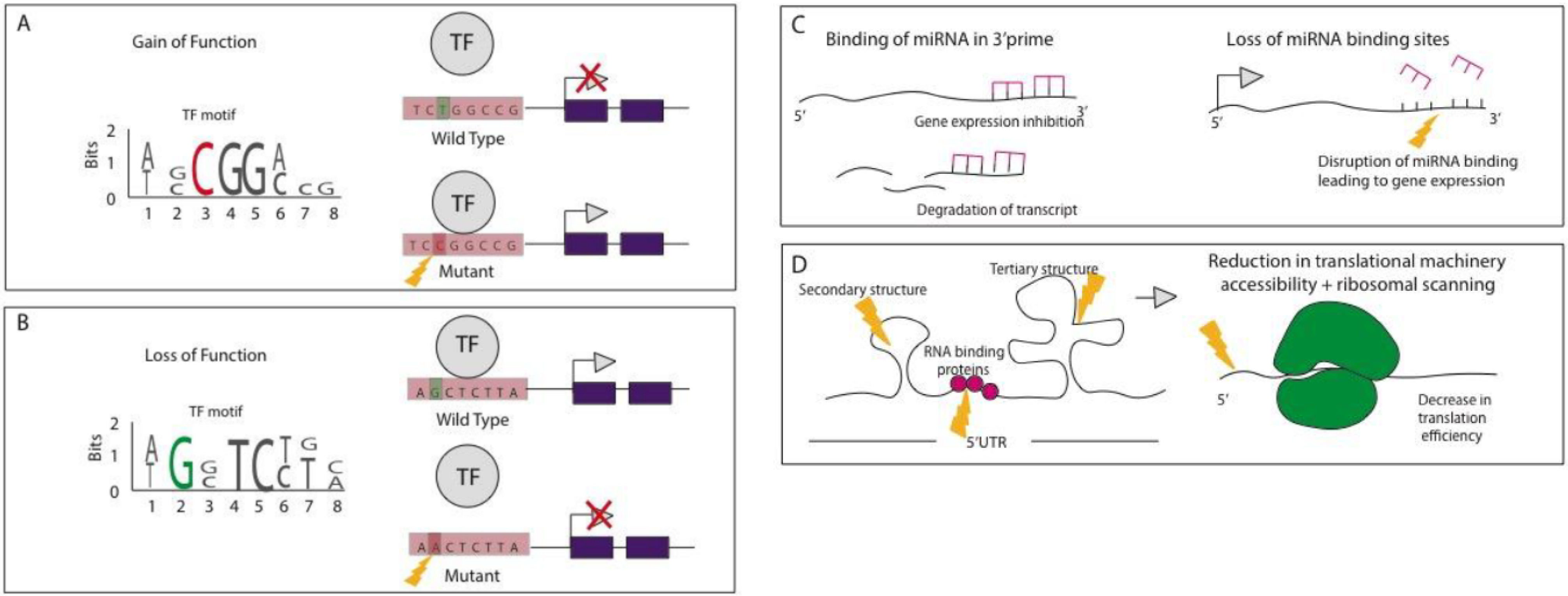
(a) Atypical chemokine receptor 1 *ACKR1 (DARC)*: mutations disrupt *GATA1* binding site $\Longrightarrow$ no expression in erythrocytes $\Longrightarrow$ no point of entry for the malarial parasite *Plasmodium vivax*

(b) Lactase *LCT*: mutations in *MCM6* intron elevate *LCT* transcription, allowing digestion of lactose

(c) Prodynorphin *PDYN*: precursor of neuropeptide dynorphin, implicated in SCZ, BP, temporal lobe epilepsy. Human-branch specific mutations (5+1) regulate constitutive and induced expression, respectively

Wray (2007) *Nat Rev Genet*

Examples of non-coding functional variants



(**A**) Mutations within promoter (e.g., *TERT*) and enhancer regions (*TAL1*) can create transcription factor (TF) binding motifs in a gain-of-function manner allowing the binding of transcriptional activators (**B**) Alternatively, mutations within regulatory regions can create the loss of transcription factor binding sites, leading to transcriptional repression (**C**) miRNA binding within the 3' UTR control gene expression, by inhibiting translation or marking transcripts for degradation. Mutations that disrupt these binding sites can lead to over-expression (*NFKBIE* and *NOTCH1* genes in cancer) (**D**) Mutations within the 5' UTR can alter the secondary and tertiary structures, as well as trans-acting RNA binding protein sites. These alterations can affect translation efficiency and mRNA stability (*BRCA1* and *CDKN2A* genes)

Patel (2018) *High-Throughput*

Examples of non-coding functional variants

The *NOS1AP* gene on human chromosome 1q has been long known to be associated with variability of **QT interval and cardiac repolarization**, whereas the underlying mechanism was unclear. A recent study utilized high-coverage resequencing and regional association for fine mapping in the GWAS locus for QT interval variation, which identified **210 common non-coding risk variants**. Further enhancer/suppressor analysis of 12 selected variants located in cardiac phenotype associated DNaseI hypersensitivity sites assisted in the identification of an upstream enhancer variant (rs7539120) associated with QT interval. This variant can affect cardiac function by increasing *NOS1AP* transcript expression in cardiomyocyte-intercalated discs and increase risk of cardiac arrhythmias.

Similar evidence for functional enhancer SNPs has also been observed at many other loci, including the intronic enhancer SNPs at the *MEIS1* gene associated with **restless legs syndrome** and at the *BCL11A* gene associated with fetal hemoglobin levels, the intergenic enhancer SNP upstream to the *MYB* gene that is a critical regulator of erythroid development and fetal hemoglobin levels, and the recessive mutations in a distal enhancer located 25 kb downstream of *PTF1A* that is associated with **isolated pancreatic agenesis**.

Zhang (2015) *Hum Mol Genet*

Examples of non-coding functional variants

A recent study on the **schizophrenia**-associated locus at 1p21.3 identified a rare enhancer SNP (chr1:98515539A>T, hg19) with increased risk. The chromatin conformation capture assay showed that this risk allele has no obvious influence on the neighboring genes such as *DPYD*, but can reduce the expression of non-coding genes MIR137/MIR2682.

In some instances, such functional variants are located in either the 5′ or 3′ untranslated region (UTR) of the disease-associated genes. A recent study identified the association of rs11603334 (a SNP located in the 5′ UTR of *ARAP1*) with **fasting proinsulin and type 2 diabetes**. The allele-specific expression assay in human pancreatic islet samples showed that the risk allele of rs11603334 can upregulate gene expression of *ARAP1* by 2-fold, which is also supported by the observation of decreased binding of pancreatic beta cell transcriptional regulators *PAX6* and *PAX4* to the rs11603334 risk allele and its corresponding increased promoter activity.

In the case of **hypertriglyceridemia**-associated *APOA5*, the 3′ UTR SNP rs2266788 was predicted to create a potential miRNA binding site for liver-expressed miR-485-5p. Luciferase reporter assays in both HEK293T cells with a miR485-5p precursor and in HuH-7 cells with endogenously expressed miR-485-5p suggested that the mutant allele of rs2266788 is involved in the miR-485-5p-mediated downregulation of *APOA5*.

Zhang (2015) *Hum Mol Genet*

Prediction of non-coding variant effect

CADD: Combined Annotation–Dependent Depletion integrates diverse genome annotations and scores *any possible* human single-nucleotide variant (SNV) or small insertion-deletion (indel) event

«Deleterious variants—that is, variants that reduce organismal fitness—are depleted by natural selection in fixed but not simulated variation»

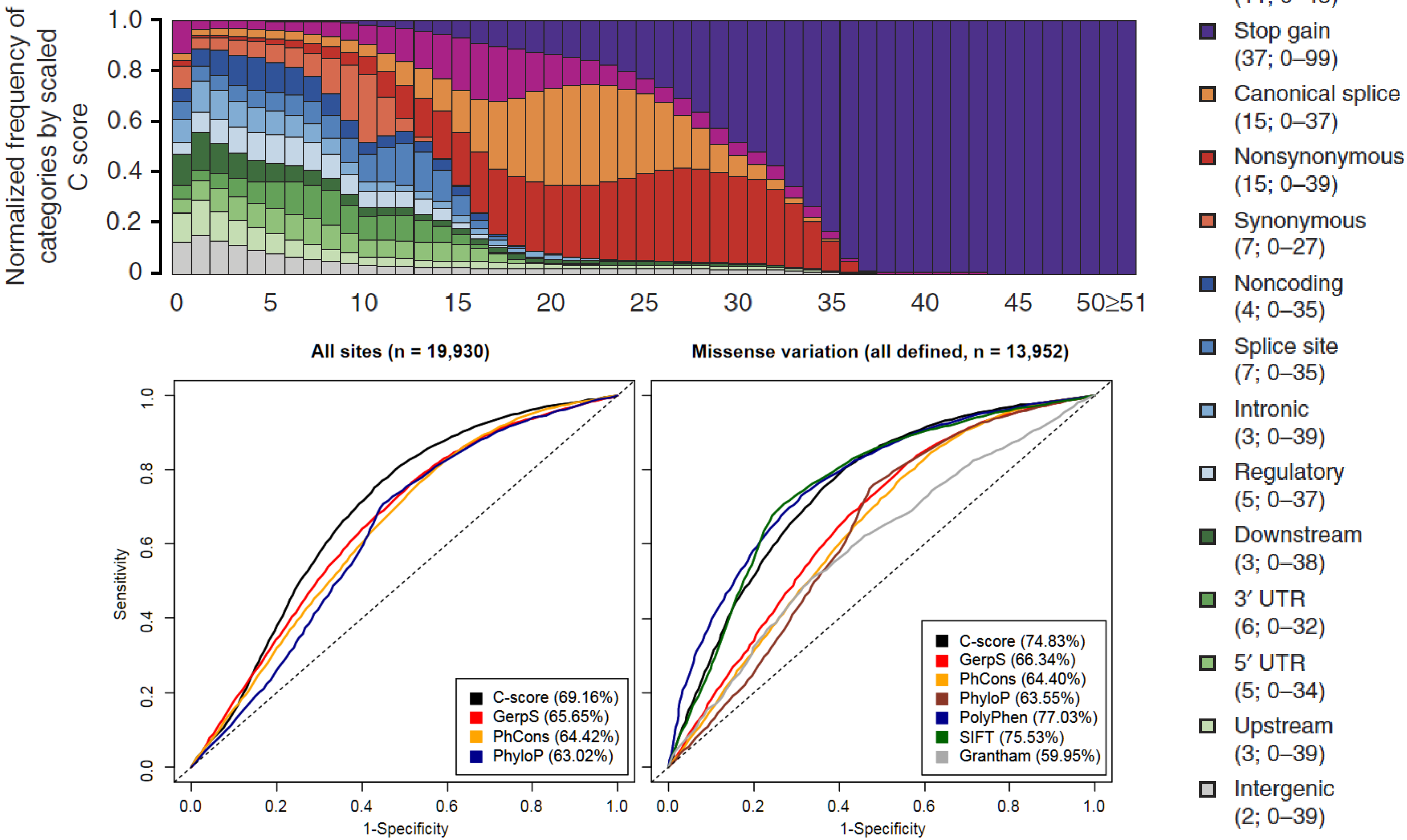
Observed variants (15 mln SNVs, 0.63 mln insertions and 1.1 mln deletions):
– human-chimp differences; SNPs with MAF>5% excluded
– SNPs with DAF (derived allele frequency) > 95% (<5% of total)
Simulated variants (44 mln SNVs, 2.1 mln insertions and 3.1 mln deletions):
– a fully empirical model of sequence evolution with a separate rate for CpG dinucleotides and local adjustment of mutation rates

Features: VEP annotation, SIFT, PolyPhen-2, conservation scores, ENCODE methylation and histone modification annotation in various cell/tissue types, TF binding sites, etc.
Output: C-scores that measure deleteriousness for 8.6×10^9 variants

Kircher (2014) *Nat Genet*

Prediction of non-coding variant effect

CADD: Combined Annotation–Dependent Depletion



ClinVar pathogenic vs population variants with matched annotation

Kircher (2014) *Nat Genet*

Prediction of non-coding variant effect

Score	Data sources	Approach
Eigen	• Uses data from the ENCODE and Roadmap Epigenomics projects	• Weighted linear combination of individual annotations • Unsupervised learning method
FunSeq2	• Inter- and Intra-species conservation • Loss- and gain-of-function events for transcription factor binding • Enhancer–gene linkage	• Weighted scoring system
LINSIGHT	• Conservation scores (phastCons, phylopP), predicted binding sites (TFBS, RNA), regional annotations (ChIP-seq, RNA-seq)	• Graphical model • Selection parameter fitting using generalized linear model based on 48 genomic features
CADD	• Ensembl variant effect predictor • Protein-level scores: Grantham, SIFT, PolyPhen • DNase hypersensitivity, TFBS, transcript information • GC content, CpG content, histone methylation	• Support vector machine
FATHMM	• 46-way sequence conservation • ChIP-seq, TFBS, DNase-seq • FAIRE, footprints, GC content	• Hidden Markov models
ReMM	• Predict potential of non-coding variant to cause a Mendelian disease if mutated • 26 features: PhastCons, PhyloP, CpG, GC, regulation annotations	• Random forest classifier
Orion	• Predict potential of non-coding variant to cause a Mendelian disease if mutated • Independent from annotation and features	• Expected and observed site-frequency spectrum of a given stretch of sequence
CDTS	• Identify constrained non-coding regions in the human genome and deleteriousness of variants • Independent from annotation and features. Uses *k*-mers	• Expected and observed site-frequency spectrum of a given heptamer

Telenti (2018) *Hum Mol Genet*

Prediction of non-coding variant effect

Table 2 Summary of genomic features used for LINSIGHT scores

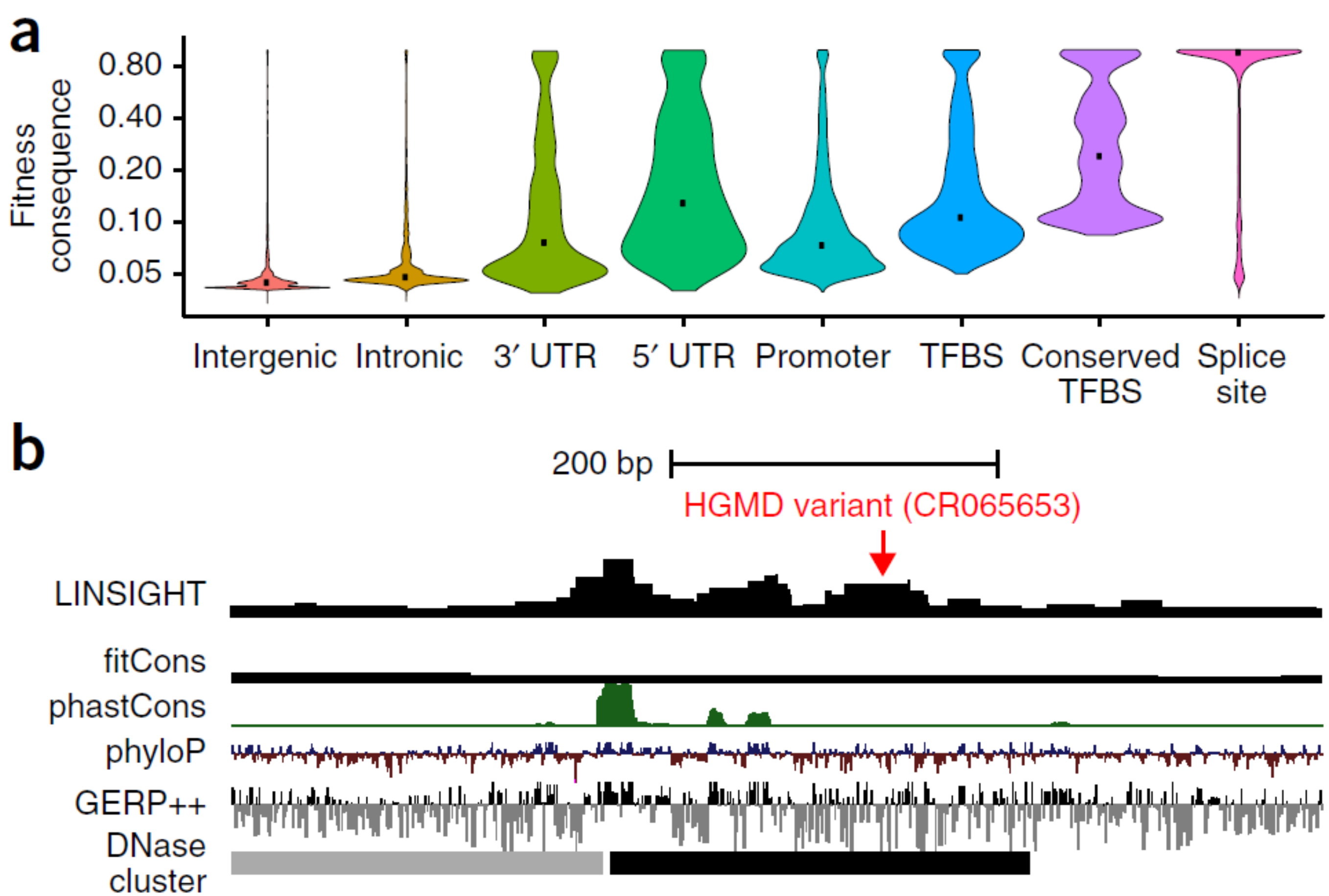
Class	Genomic feature[a]	Spatial resolution
Conservation	phyloP score	High
	phastCons element	High
	SiPhy element	High
	CEGA element	High
Binding site	Conserved TFBS	High
	rVISTA TFBS	High
	SwissRegulon TFBS	High
	Predicted TFBS within ChIP-seq peak	High
	Conserved miRNA binding site	High
	Splicing site predicted by SPIDEX	High
Regional annotation	ChIP-seq peak of transcription factor	Low
	DNase-I hypersensitive site	Low
	UCSC FAIRE peak	Low
	RNA-seq signal	Low
	Histone modification peak	Low
	FANTOM5 enhancer	Low
	Predicted distal regulatory module	Low
	Distance to nearest TSS	Low

[a]Each 'genomic feature' listed here may actually correspond to multiple features in the model. For example, four features are derived from phyloP scores: two from the mammalian phyloP scores and two from the vertebrate phyloP scores. See **Supplementary Table 3** for complete details.

LINSIGHT integrates functional genomic data together with conservation scores and other features to provide a high-powered, high-resolution measure of potential function.

Huang (2017) *Nat Genet*

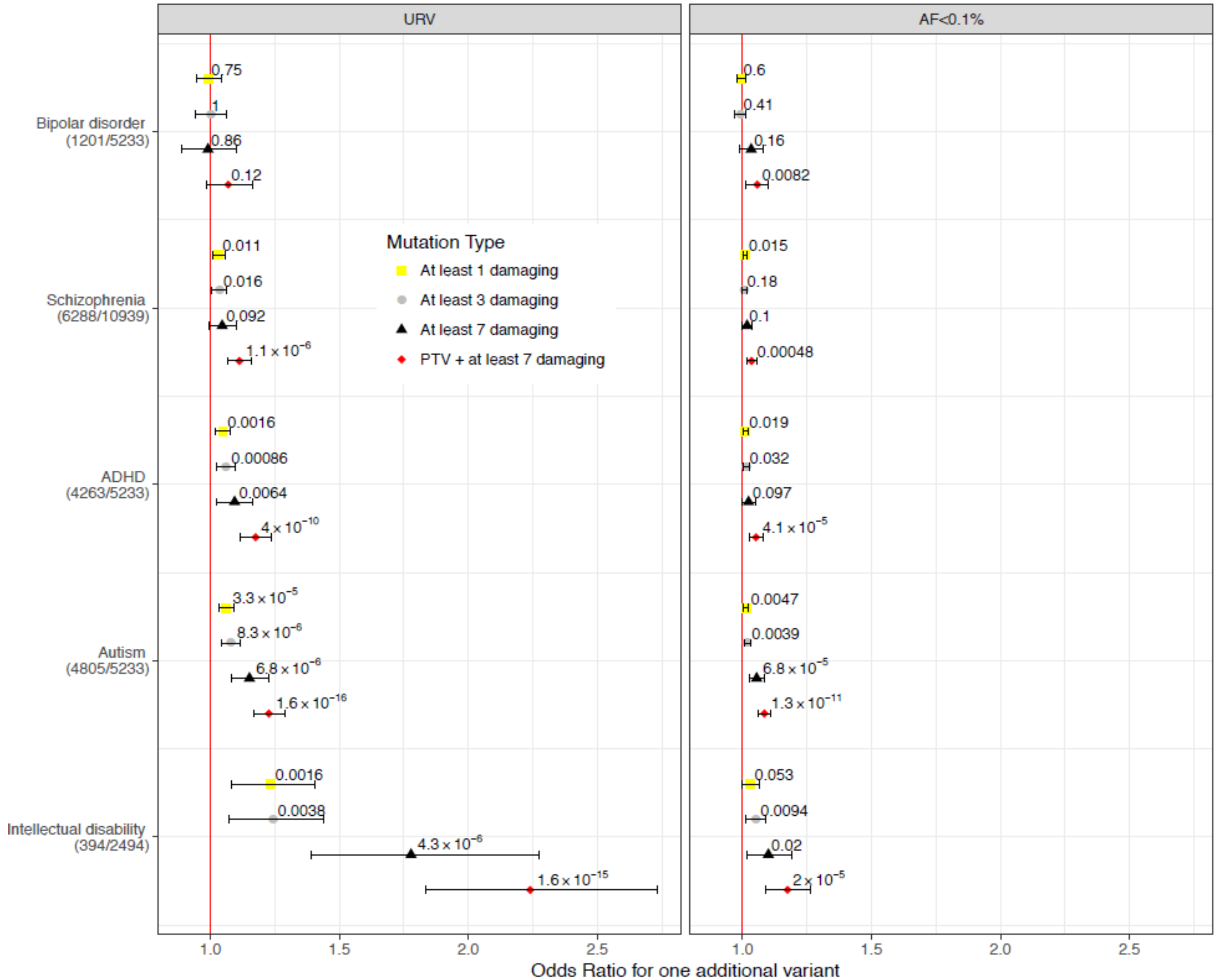
Prediction of non-coding variant effect



(**a**) Distributions of LINSIGHT scores for various genomic regions. Intergenic regions, intronic regions, UTRs, and 1-kb promoters: GENCODE 19; TFBSs: ChIP-seq peaks (Ensembl Regulatory Build); conserved TFBSs: UCSC Genome Browser. (**b**) LINSIGHT is the only method to highlight a variant from HGMD (CR065653) that is associated with upregulation of the *TERT* gene.

Huang (2017) *Nat Genet*

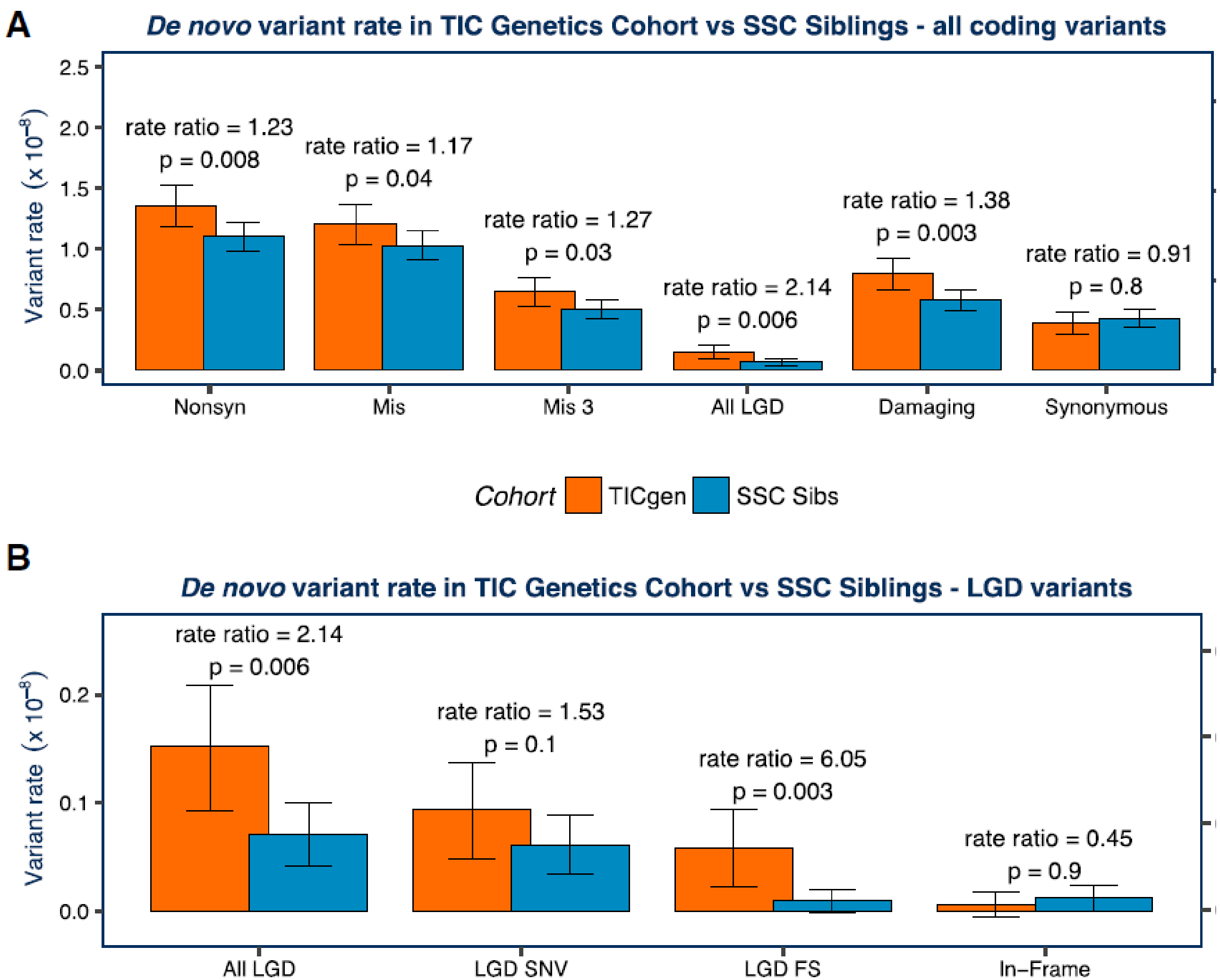
Variant effect and association with phenotypes



Meta-analyzed association between ultra-rare and rare damaging missense variants in PTV-intolerant genes and 5 diseases. **The strength of the association increases as function of the number of algorithms and is particularly strong among ultra-rare variants**

Ganna (2018) *Am J Hum Genet*

Variant effect and association with phenotypes



All classes of *de novo* non-synonymous variants show a higher mutation rate in Tourette disorder probands (orange) versus SSC siblings (controls, blue). LGD: likely gene disrupting variants: insertion of premature stop codon, frameshift, or canonical splice-site variant; **FS**: frameshift indels; **Damaging**: variants predicted by PolyPhen2; **Mis3**: LGD or damaging; **Nonsyn**: missense or nonsense

Willsey (2017) *Neuron*

Summary

- Human genome sequence is still being updated. We may soon switch from a single reference sequence to multiple ones
- Protein-coding genes represent only a minor fraction of all human genes and a tiny fraction of the genome
- Roughly one half of human genome are repetitive sequences
- Human gene structure and processing is quite diverse and complicated
- There are multiple sequence regions that assist in gene splicing: exonic and intronic splicing enhancers and silencers. A significant fraction of human disease mutations are believed to be splicing-related
- Epigenetics provide heritable phenotype changes that do not involve alterations in the DNA sequence: DNA methylation at CpG nucleotides, covalen modification of histone proteins. Noncoding RNAs are considered as part of epigenetic machinery.

Summary

- Approximately 100 genes on various chromosomes are subject to chromosomal imprinting
- Variant annotation is a procedure that determines variant consequence for a gene/protein based on its location relative to the gene sequence. It is governed and complicated by transcript structure complexity.
- Variant effect prediction determines potential functional impact of a particular variant based on its features.
- There are numerous prediction algorithms for major types of variants. Their performance and domain of applicability is a debated question, however, phenotype-associated variants are typically enriched with functional predictions.

Further reading

- Strachan, Read – *Human Molecular Genetics*, Chapter 13
- Rivas, M.A., Pirinen, M., Conrad, D.F., Lek, M., et al. (2015). Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* 348, 666–669.
- Saleheen, D., Natarajan, P., et al. (2017). Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* 544, 235–239
- Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 176, 535-548.e24.
- Niroula, A., and Vihinen, M. (2016). Variation Interpretation Predictors: Principles, Types, Performance, and Choice. *Human Mutation* 37, 579–597.

Further reading

- Li, J., Zhao, T., Zhang, Y., Zhang, K., Shi, L., Chen, Y., Wang, X., and Sun, Z. (2018). Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic Acids Res* 46, 7793–7804.

- DePristo, M.A., Weinreich, D.M., and Hartl, D.L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet* 6, 678–687.

- Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am. J. Hum. Genet.* 102, 11–26.

- Lee, P., Lee, C., Li, X., Wee, B., Dwivedi, T., and Daly, M. (2018). Principles and methods of in-silico prioritization of non-coding regulatory variants. *Hum Genet* 137, 15–30.

- Eilbeck, K., Quinlan, A., and Yandell, M. (2017). Settling the score: variant prioritization and Mendelian disease. *Nature Reviews Genetics* 18, 599.